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

December 12, 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. I've spent this past summer and fall working in computational biology at the Christina Leslie Lab at Sloan Kettering Institute, 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.

I've been working on designing a statistically inspired robust pipeline for analysis of 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 function better.

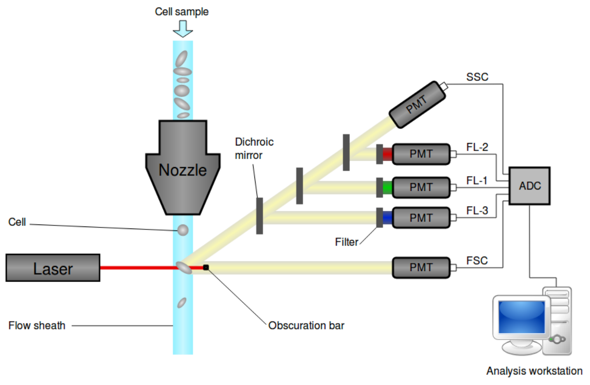


Figure 1: Structure and working of a flow cytometer

2 Analysis of Flow Cytometry Data

2.1 Flow Cytometry

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 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).

2.2 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 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.

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 such arbitrary thresholds lead to unstable classification which isn't reproducible. Moreover, 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, robust and statistically inspired approach at dividing the sample into various known populations.

2.3 Proposed Analysis Pipeline

The first pipeline we tried focused on the use of Louvain clustering with a tsne visualization for quantifying the immune cells population - refer to Fig. 3 for a description of white blood

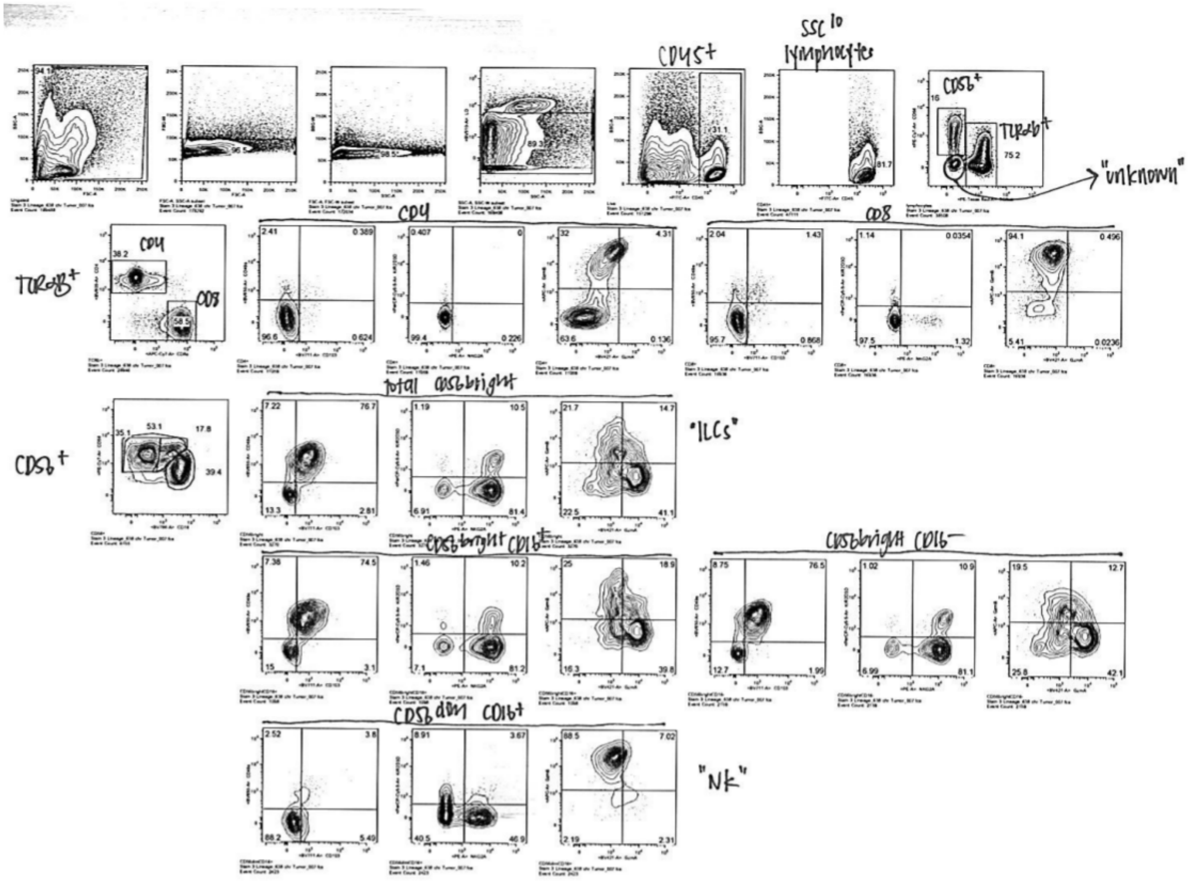


Figure 2: Manual gating workflow: tree like gating structure is followed to identify populations at various granularities

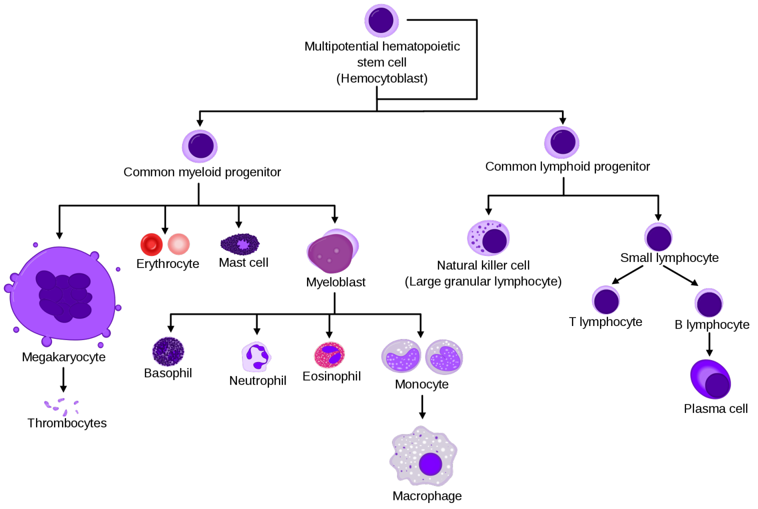


Figure 3: White blood cell lineage

cell lineage. The proposed pipeline for analysis is:

- Use the compensated data with preliminary filtering for live, CD45+ singlets
- Perform a logicle transform on the data. This is a slight modification on the biexponential transform to scale the fluorescence data for classification
- Perform k-Nearest Neighbours (kNN) and then louvain clustering on the nearest neighbour set graph
- Visualise the high-dimensional data using t-SNE embedding
- The phenotype 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.

We then started working on trying to generalize this to other patients and other panels, and we started running into problems. Due to the inherent nature of the nearest-neighbour classification, increasing resolution often breaks up clusters that are biologically the same, thus creating structure where there isn't. This is a downside of a completely global approach.

So, we changed our efforts into looking at Gaussian Mixture Models to aid us. GMMs are known to work pretty well in low dimensions, but the data to be clustered here is high-dimensional (11D) and a vanilla GMM would easily overfit the data, as a 11x11 covariance matrix needs a huge amount of data to learn (curse of dimensionality). Instead, we hypothesized that the covariance of each cluster actually lives in a more low-rank space, with most markers being independent of each other. With this in mind, we tried using a 1D GMM successively on clearly bimodal markers to classify them into populations recursively, till we get populations that are clearly unimodal in all markers. This helps us estimate the number of clusters and their initialization means (terminal nodes in gating tree). We then use this information along with our hypothesis to train a Sparse GMM (sGMM), by enforcing the trained covariance matrices to mostly consist of zeros. This is achieved by penalizing the likelihood function with an L1 regularization, achieved using the graphical lasso method. This method is still in the works and is being tweaked to be robust and avoid convrgence issues, but it's already proving more scalable than louvain clustering.

3 Results

For the tumor sample of a chromophobe RCC patient with a lineage panel, the results are shown in Fig. 4 & 5, providing us with insight into what these individual populations are: 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

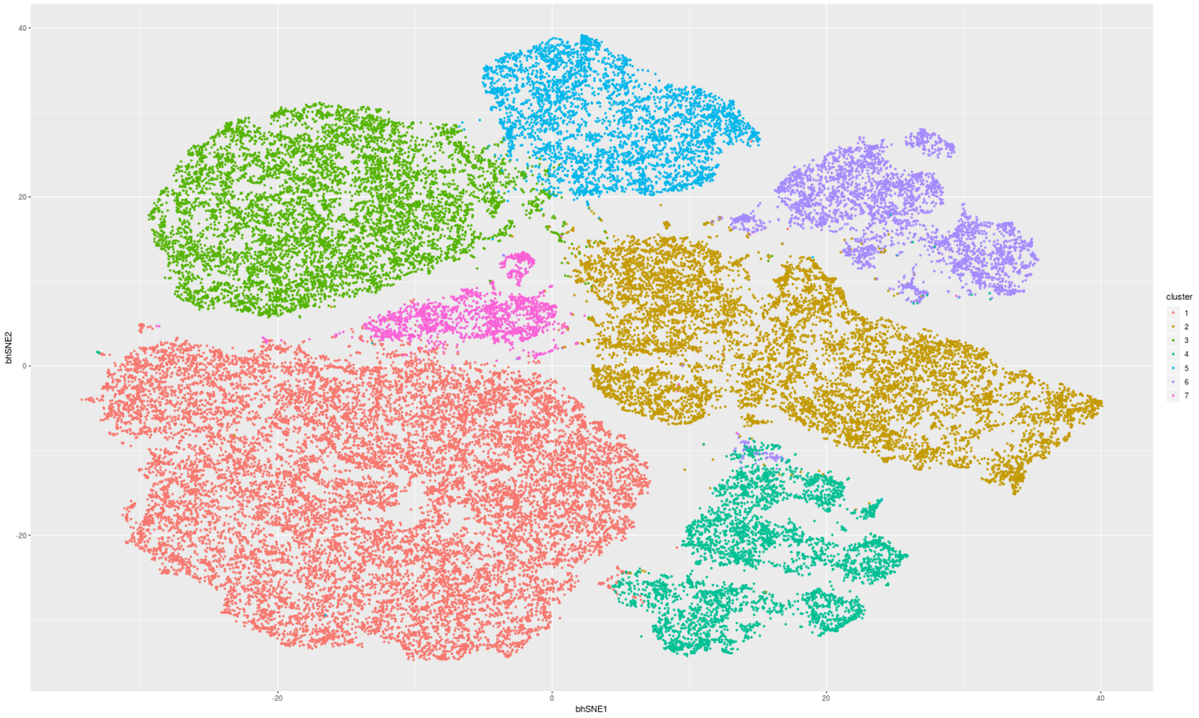


Figure 4: Louvain clustering/t-SNE with 7 clusters

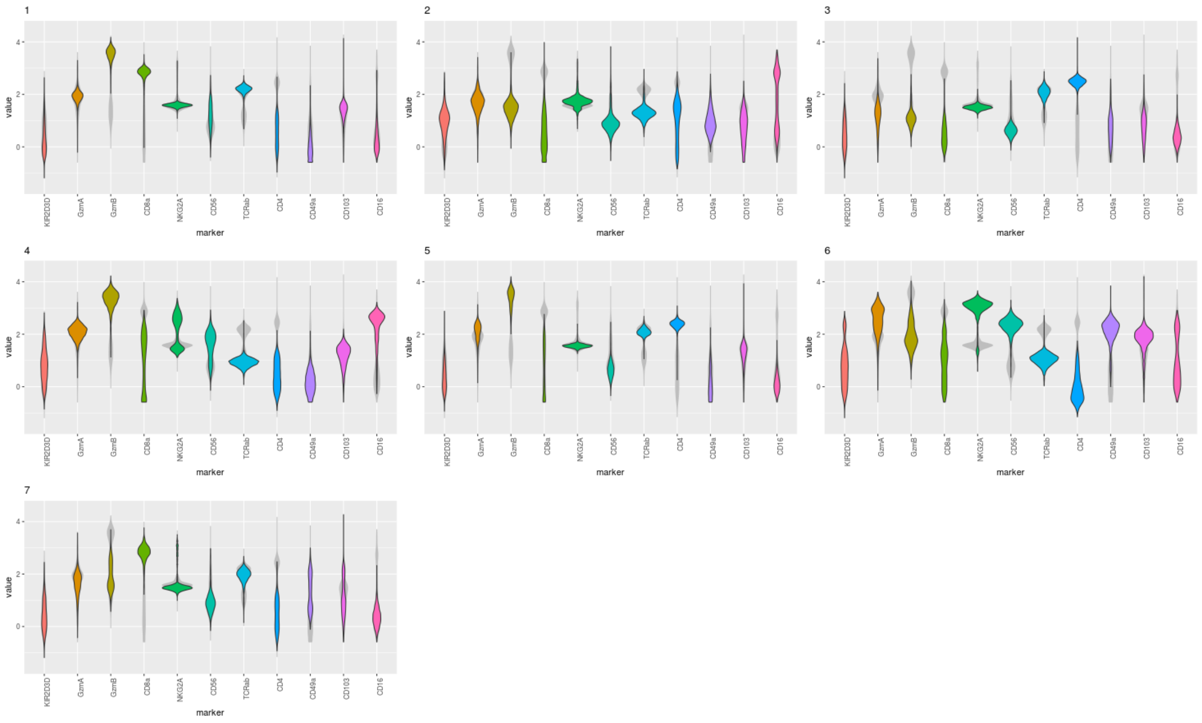


Figure 5: Violin plots for protein marker expression of each cluster

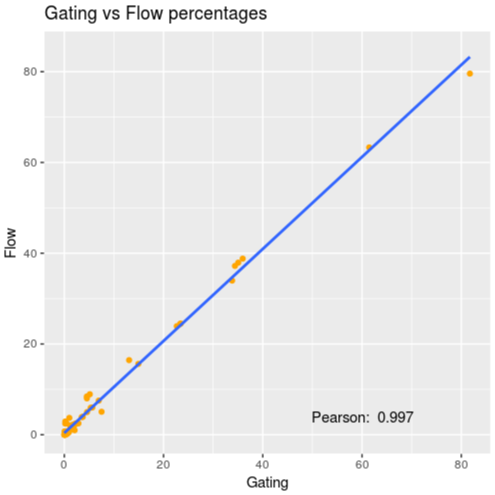
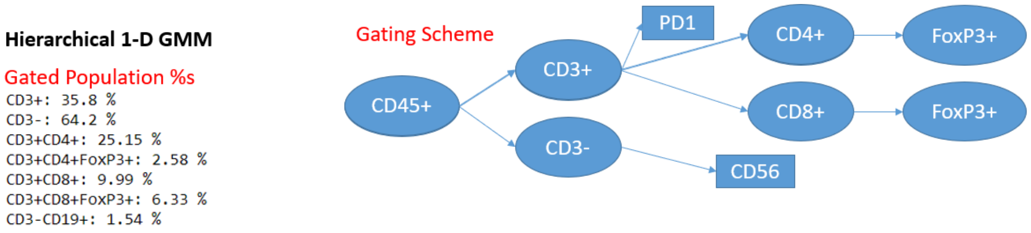


Figure 6: Correlation between manual gating and clustering percentages

resident (CD49a+,CD103+) and thus consists of Innate Lymphocytes (ILCs). Cluster 2 is what we 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.

For the Gaussian Mixture Model method, in a T-cell panel, following along a typical “statistical” gating scheme as in Fig. 7, we perform 1D GMM at each tree node to get various populations, shown in Fig. 8.



```
CD3+: 35.8 %
CD3-: 64.2 %
CD3+CD4+: 25.15 %
CD3+CD4+FoxP3+: 2.58 %
CD3+CD8+: 9.99 %
CD3+CD8+FoxP3+: 6.33 %
CD3-CD19+: 1.54 %
```

Figure 7: Gating tree and node percentages and for Hierarchical 1D-GMM

4 Discussion & Comparison

In essence, the difference between classical flow analysis and the louvain clustering 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 cluster things statistically. With manual gating, we can have a more precise

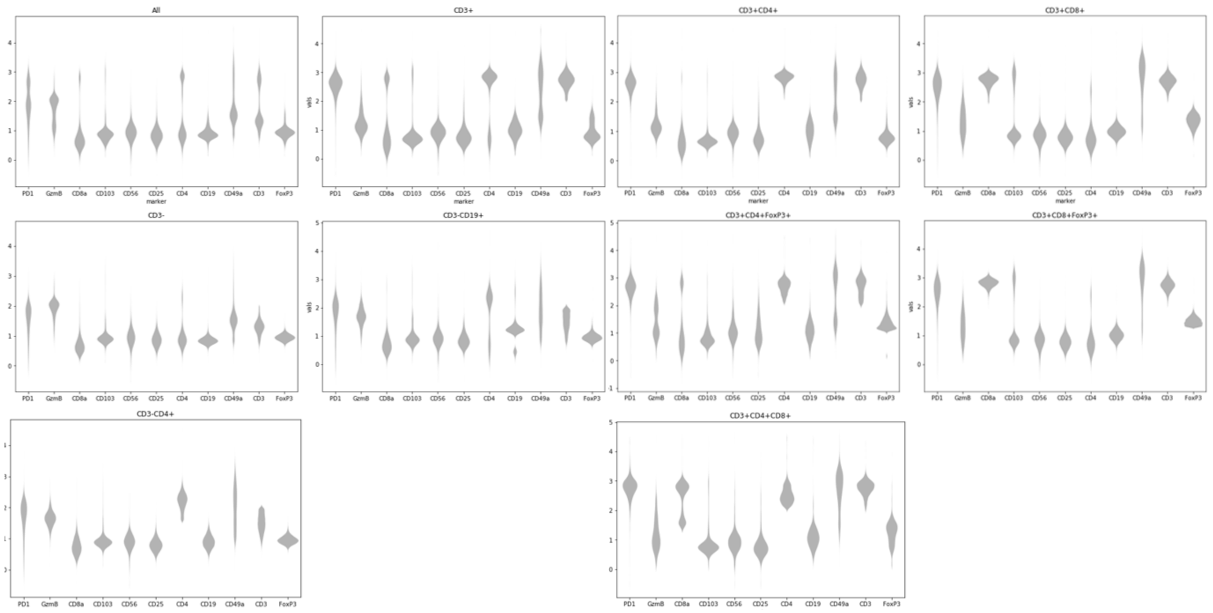


Figure 8: Marker distributions at various levels of gating tree

look at marker activities and divide the populations based on biological plausibility, but the thresholds are arbitrary and consecutive gates on 2D slices can tunnel-vision us and make us miss out on populations. With the proposed louvain 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 multidimensional statistics, this might give us more sensible divisions. However, the introduction of meaningless clusters at higher resolutions drives us to a method that's at least partially biologically inspired from normal gating practices. To overcome the shortcomings of both of these and incorporating their strengths, we propose using statistical gating using GMMs, followed by using the cluster number and mean initializations in a sGMM framework, which would lead to more biologically meaningful populations.

4.1 Future Work

I'll be joining the lab as a full-time Research Assistant in Feb 2020. The future work on this flow project would be to follow along the sGMM analysis and comparing with louvain clustering. Other avenues to explore are modeling batch effect and inter-patient variation, and use of scRNA-seq in sync with flow data for comprehensive analysis. If we are successful in our efforts, it would be really cool, as it would allow us to rescue all this RCC flow data that has been collected over two years. It would not only augment the current scRNA-seq analysis, but might prove to be a novel method for a more end-to-end, more robust and reproducible method of quantifying the immune cell population using statistics and machine learning in high-dimensional data.