Outline – Automated Flow Cytometry Data Analysis Pipeline with OpenCyto

John Ramey, Greg Finak, Mike Jiang, Raphael Gottardo, Jafar Taghiyar, Nima Aghaeepour, and Ryan Brinkman

January 7, 2013

Hypothesis: Automated gating can perform as well or better than manual gating to identify responding T-cell subpopulations in vaccine clinical trial data.

1. Introduction

   - Goal: Use separation of meaningful cell subpopulations identified via automated gating (pre- and post-vaccination) to ~~prognosticate~~ identify subjects with a vaccine response.
   - We use two intracellular cytokine staining (ICS) data sets produced by the HIV Vaccine Trials Network (HVTN).
   - Emphasize the pitfalls of manual gating; automation is preferred.
     manual gating is time-consuming, and may be subjective if experiments are not well controlled.
   - OpenCyto can recapitulate manual gating via an automated pipeline that incorporates prior knowledge through a Bayesian framework.
   - OpenCyto provides fast, robust automated gating of large data using a manual gating strategy specified by the user.

2. OpenCyto

   - Discuss infrastructure
   - Describe the data-analysis pipeline
   - Describe the different gating approaches
   - Emphasize that gating is data-driven and can incorporate expert opinion as well as marker-specific, data-driven priors
   - Gating is performed in one and two dimensions, so that the gating results are easy to understand

2.5 Show that OpenCyto can be used to replicate manual gating with respect to low variability and bias compared to the manual gating statistics.

Identify responders in the traditional way (stimulated vs unstimulated within each time point).

3. Describe classification details

   Show that we can improve discrimination by using more features and optimizing the gating, and making this a classification problem (when baseline is available).

   - Optimize gating set for classification of pre- and post-vaccination individuals

     Rapidly prototype different gating thresholds for the cytokines.

   - Describe how our classifier is constructed

     i.e. You could design the pipeline for the ENV stimulation and run it on the GAG stimulation for each data set to show that it is robust when data is standardized

     - We extract all Boolean subsets with associated proportions as features
     - Briefly provide example Boolean subset, similar to FlowCAP 3 talk: (CD4) IL2+ and !IFNg+ and TNFa+
     - We then utilize a LASSO-based classifier using the `glmnet` R package
     - Mention briefly that the shrinkage parameter selected via cross-validation
     - Mention also that `glmnet` employs a variable selection via L_1 regularization

4. Discuss data sets and results

   - Data sets
     - Data set #1: HVTN 065
     - Data set #2: HVTN 080
   - Results
     - Emphasize that OpenCyto yields similar results to the manual gating
     - Discuss the features selected by `glmnet`
     - Figures:
       - Figure 1: Output from flowClust that demonstrates the fitted mixture model
       - Figure 2: Comparison of automated and manual gates
       - Figure 3: Gated proportions of stimulation groups by features selected for each training subject

       Figure: comparison of manual gating and automated gating statistics. (show CVs)

       Table: Response rates by treatment group for each data set (manual and automtaed).

5. Discussion