

Analyzing ICS Assays Using a BioConductor Pipeline

Greg Finak, Mike Jiang

Gottardo Lab

Fred Hutchinson Cancer Research Center

Highlights

- Our interest was in demonstrating the utility of simple, automated flow analysis tools in BioConductor.
- Pipeline uses only core BioConductor toolset.
- Knowledge-driven gating strategy mimics manual analysis.
- Methodology is fast, reproducible, and easy to interpret.
- Difficulty: dealing with rare populations.

Outline

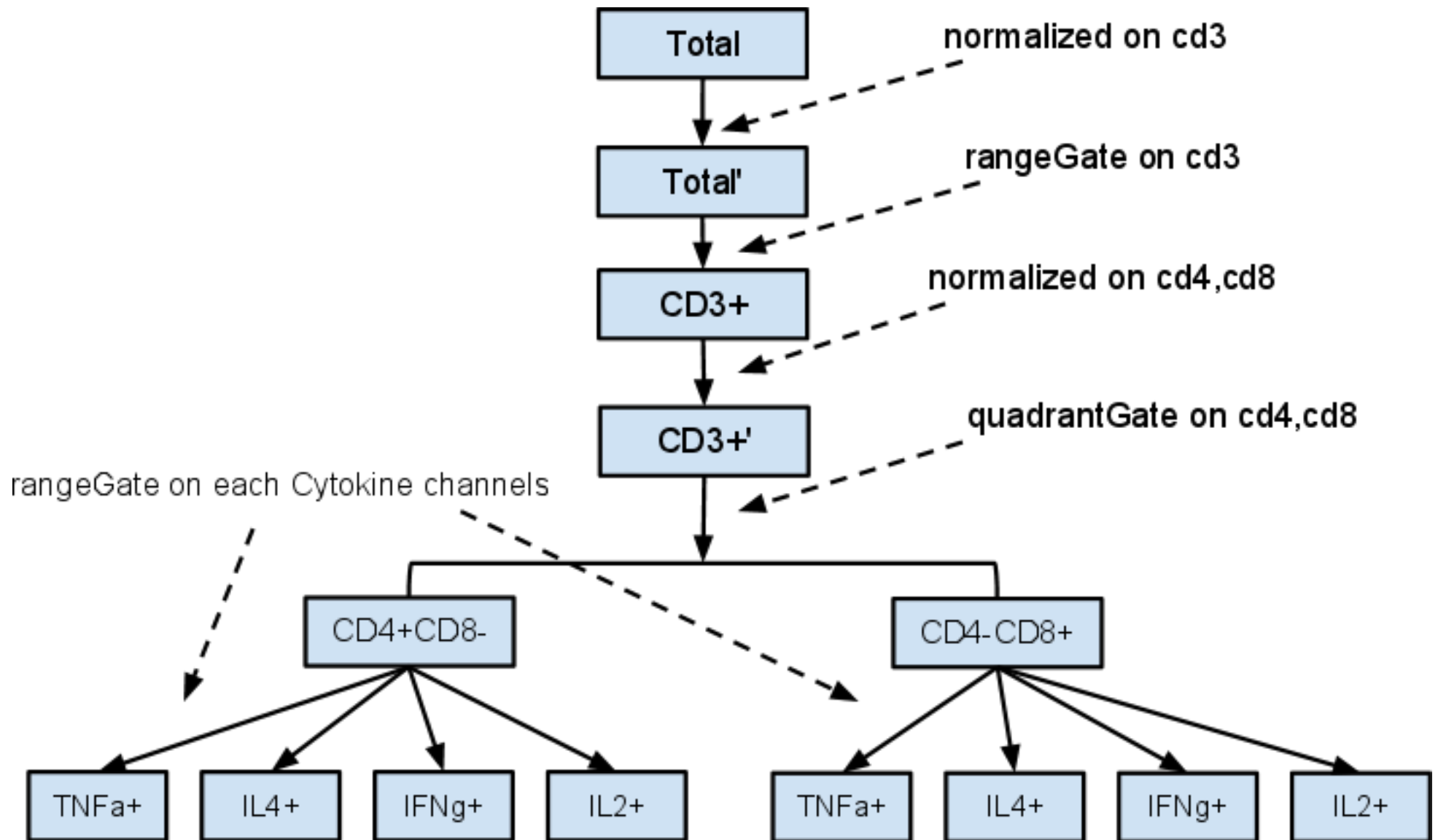
- **Preprocessing and Gating**
 - Sequential normalization
 - Gating strategy
- **Challenge 3a (ENV / GAG classification)**

- **Challenge 3b (Responder / Non-Responder Calls)**

Data Description

- 48 individuals (240 FCS files)
 - 5 FCS files for each individual :
 - 2 Stimulations (ENV/GAG)
 - 2 negative controls and 1 positive control
 - Training: 27 subjects (135 FCS)
 - Testing: 21 subjects (105 FCS)
- compensated, transformed and partially gated (for singlets, live cells and lymphocytes).
- Markers
 - CD3/CD4/CD8
 - TNFa/IL4/IFNg/IL2
- Goals
 - Classify the **antigen stimulation**
 - Classify each sample as either a **responder** or **non-responder**

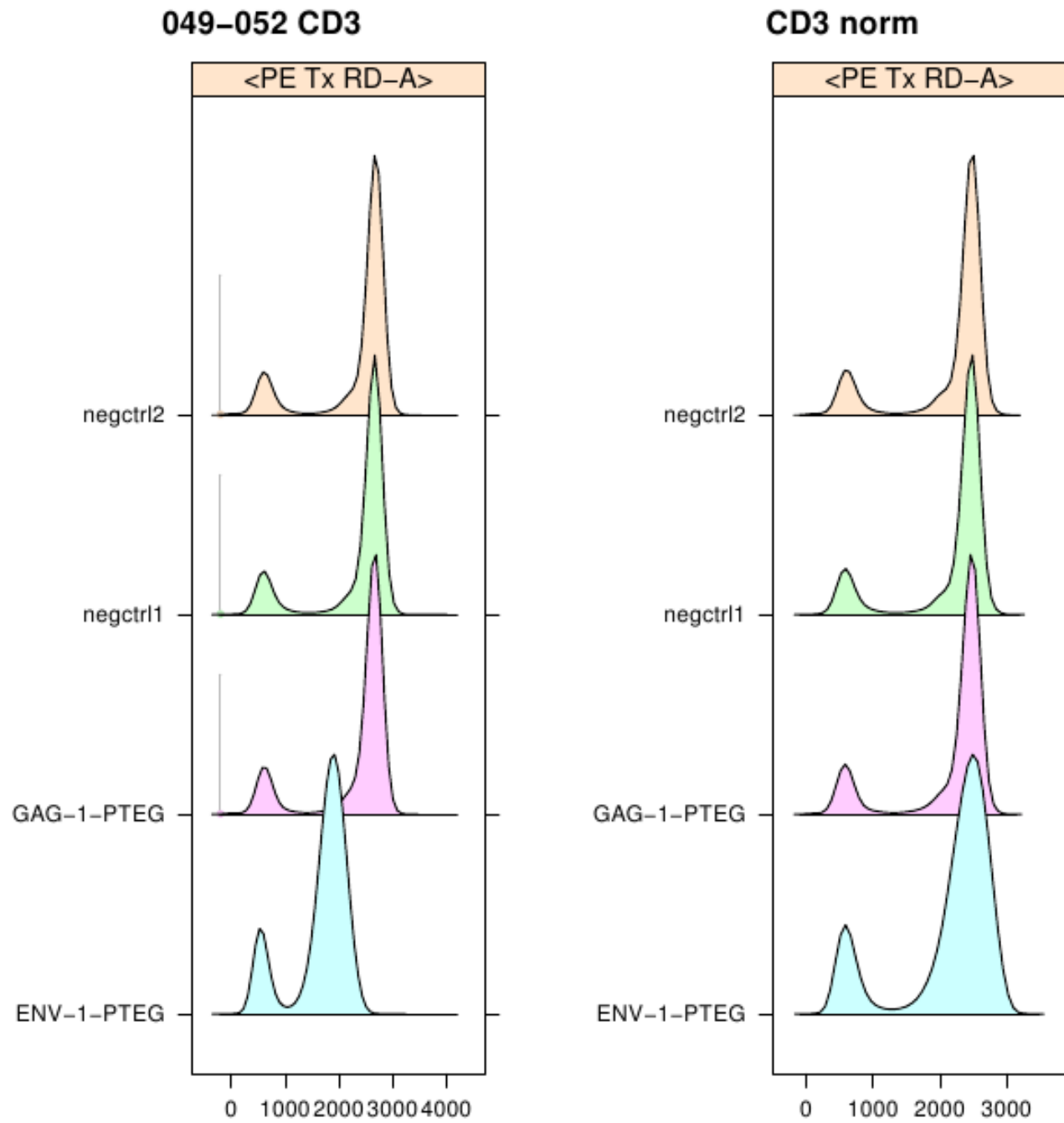
Sequential normalization/Gating



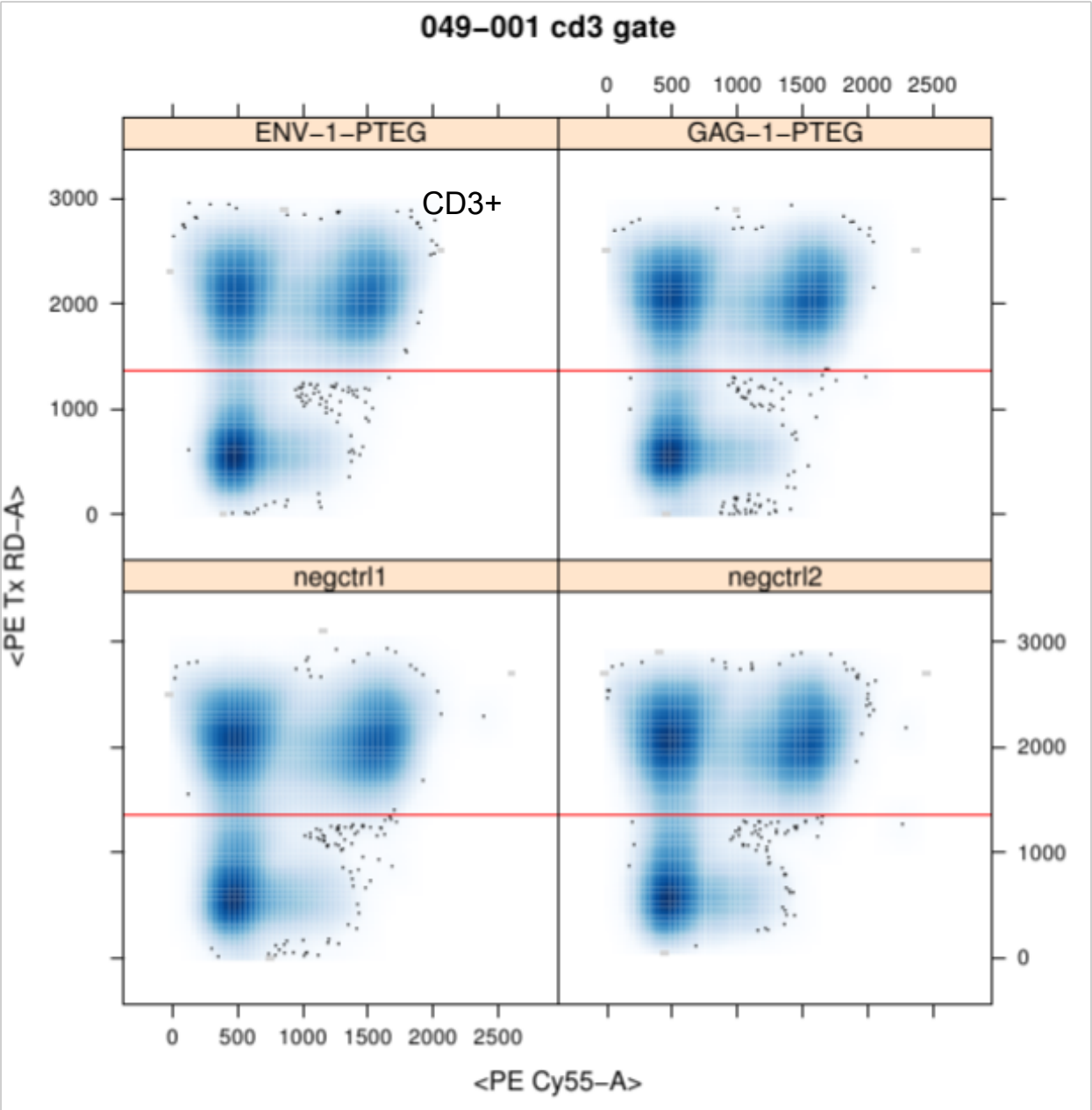
BioConductor Tools

- **flowCore (updated)**
 - Core functionality for flow cytometry data analysis in BioConductor (flowSets, flowFrames).
- **flowStats (updated)**
 - Convenience methods for 1D, 2D gating (rangeGate and quadGate)
 - flow cytometry data normalization (warpSet, warpSetNCDF, warpSetNCDFLowMem)
- **ncdfFlow (new)**
 - netCDF (disk-based) analysis of large flow cytometry data sets.
 - All **flowCore** functionality on netCDF backed flow data in R.
- **flowWorkspace (new)**
 - Import flowJo workspaces into R/BioConductor. Used to prepare and distribute challenge 3 data for flowCAP II.

Normalization on CD3



RangeGate on CD3 channel



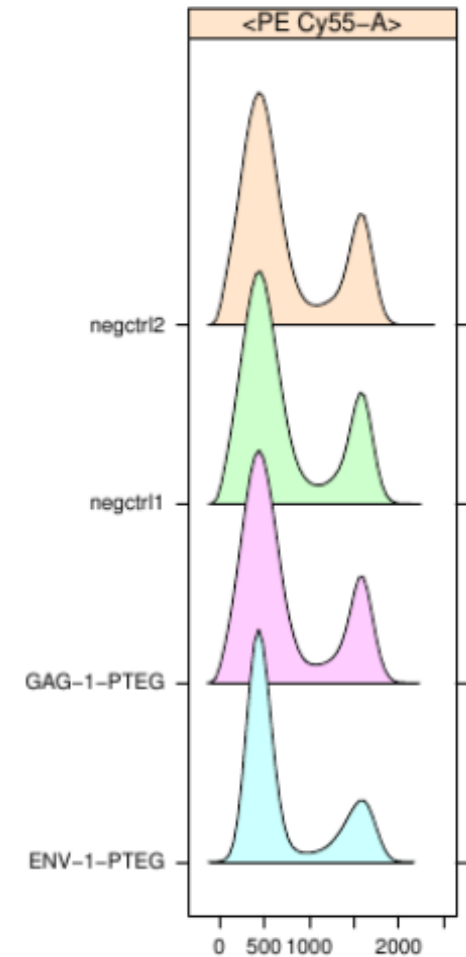
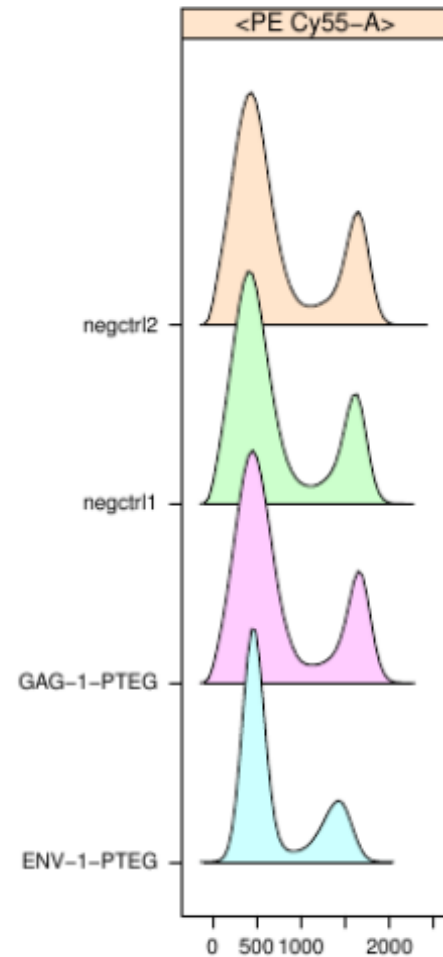
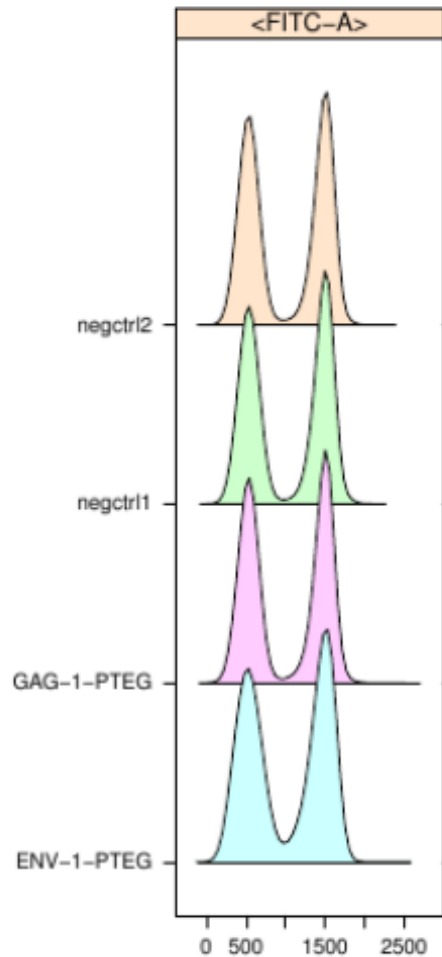
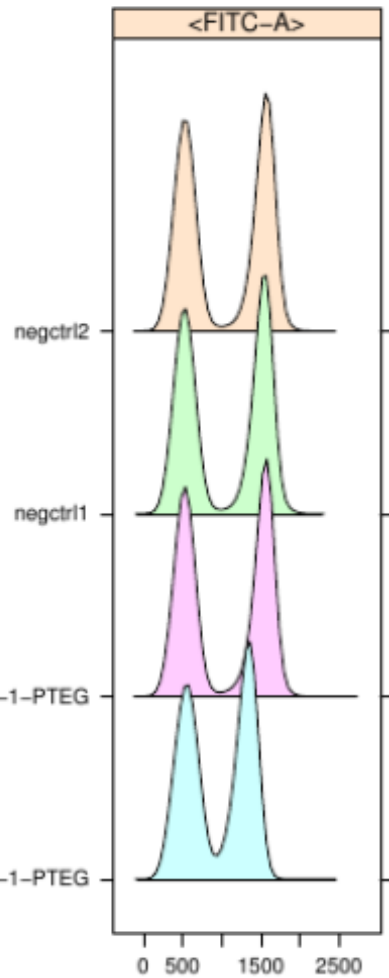
Normalization on CD4,CD8

049-052 CD4

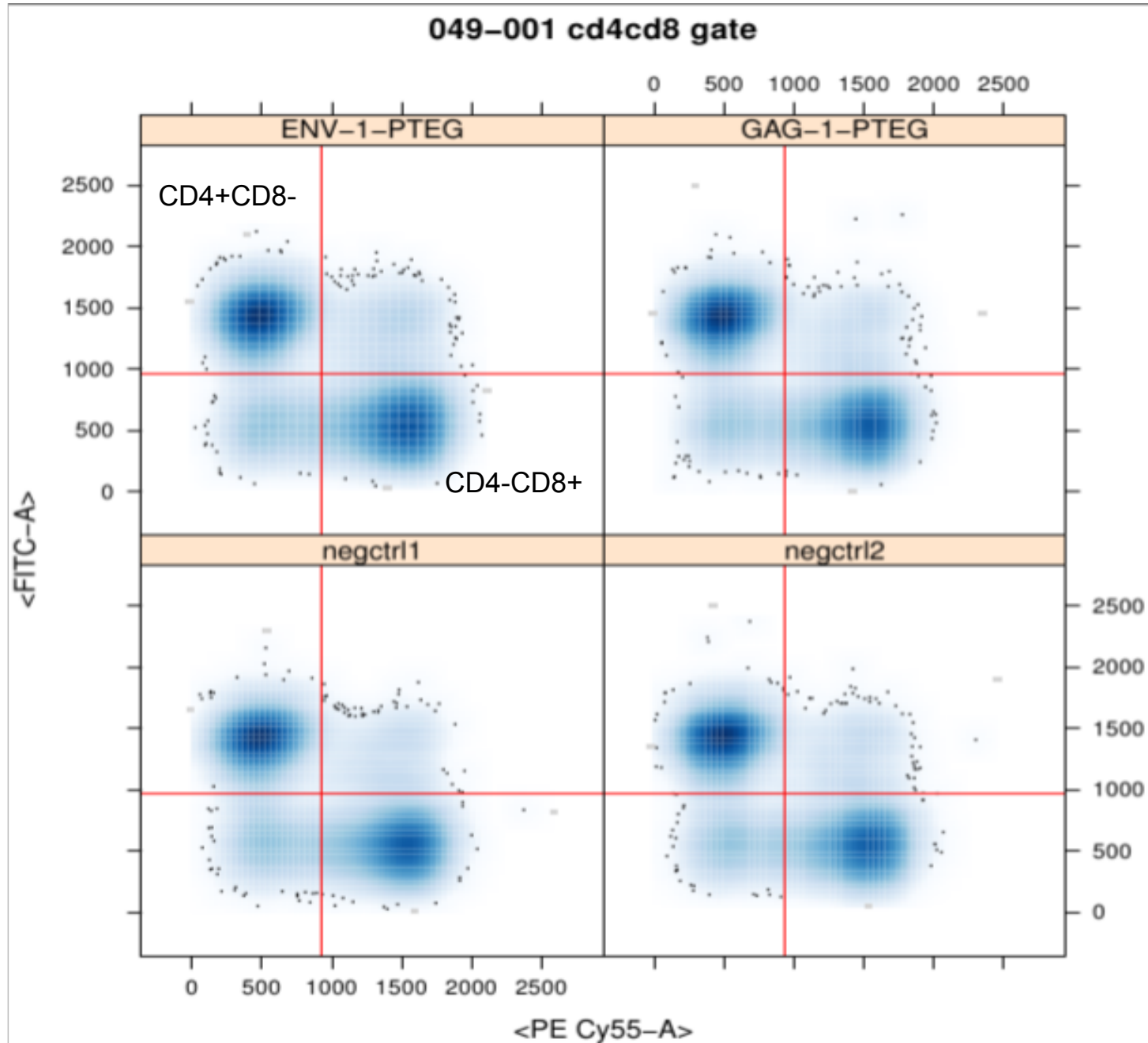
CD4 norm

CD8

CD8 norm

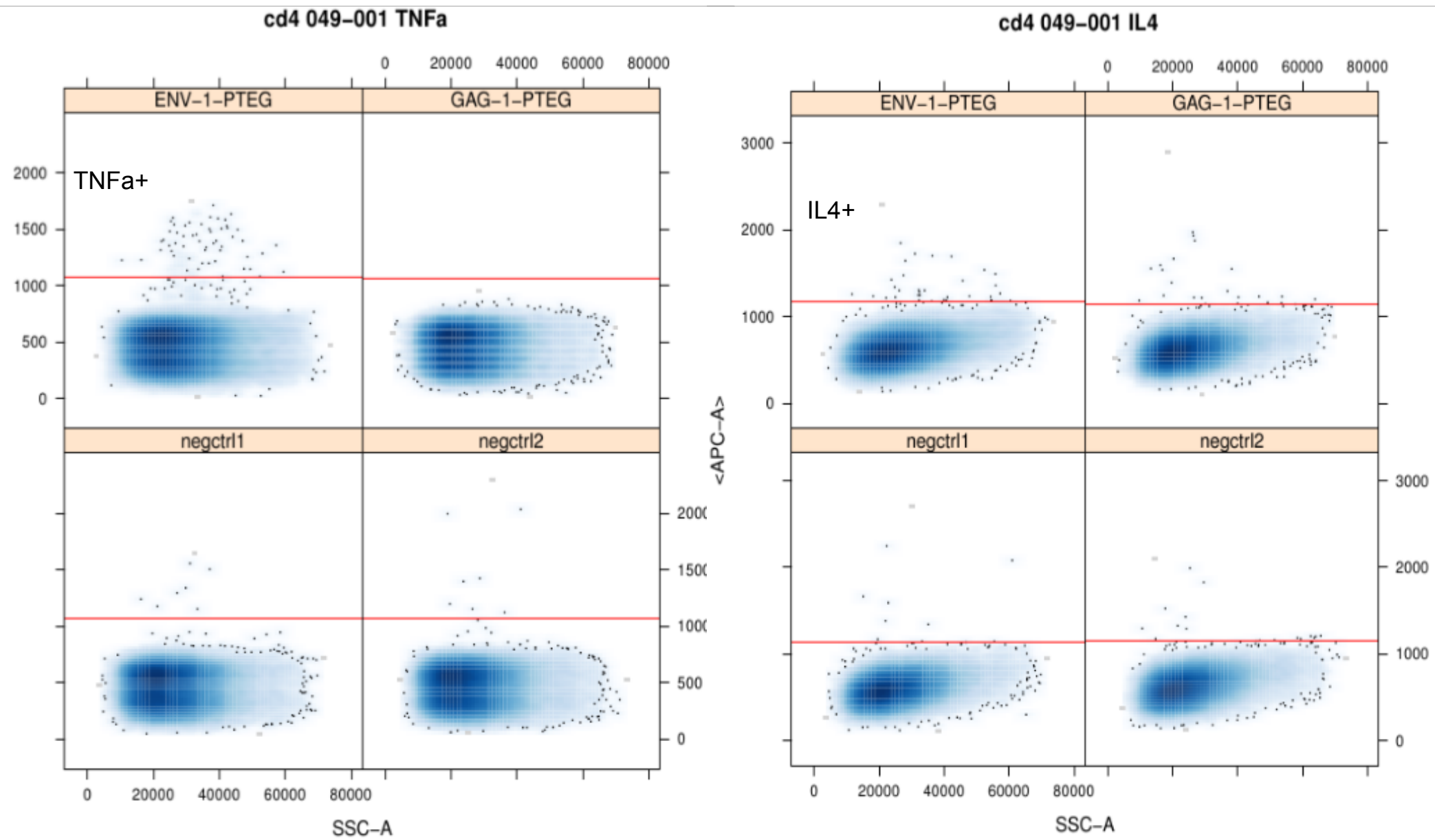


QuadrantGate on CD4 vs CD8

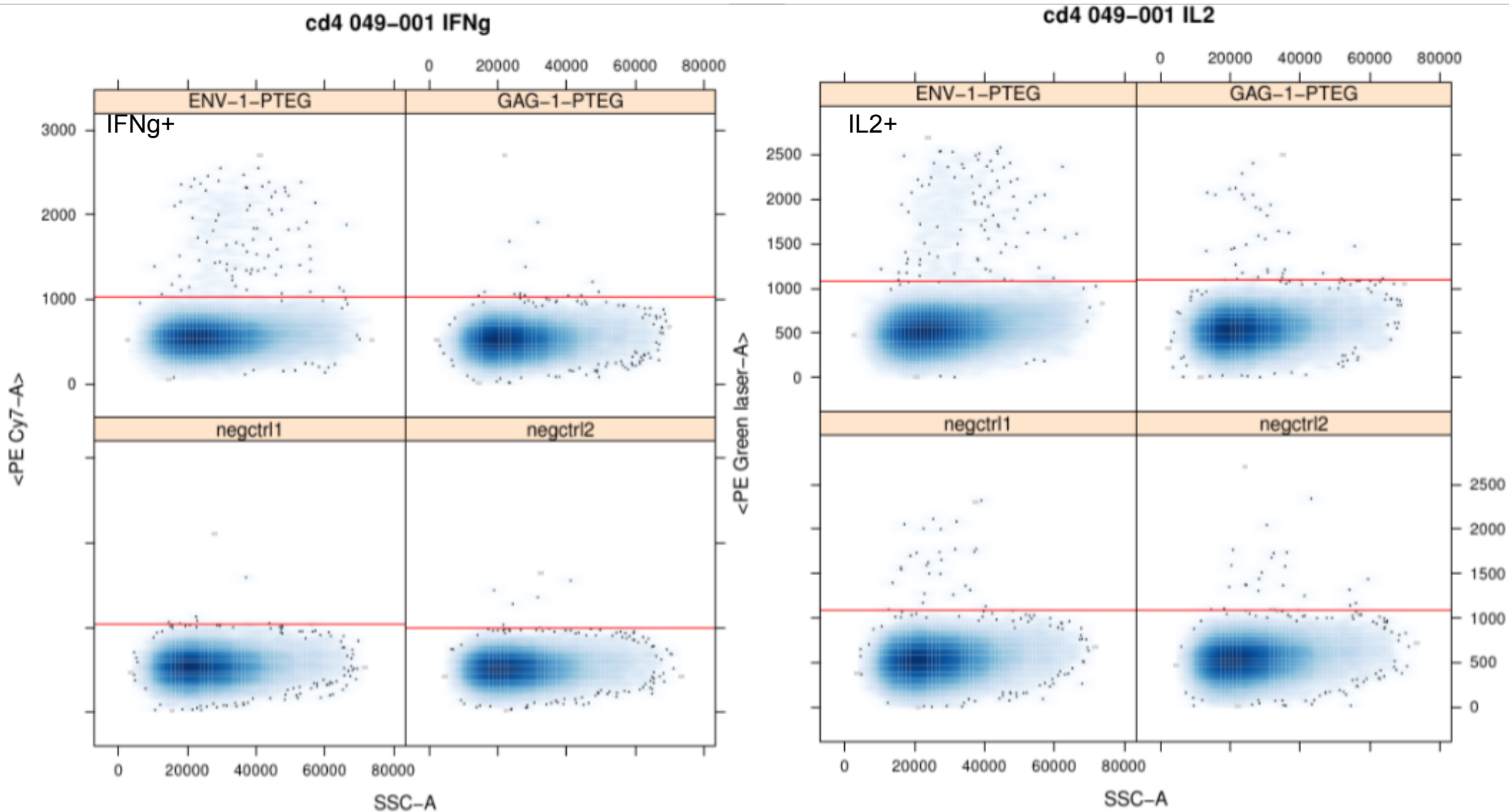


RangeGate on Cytokine channels

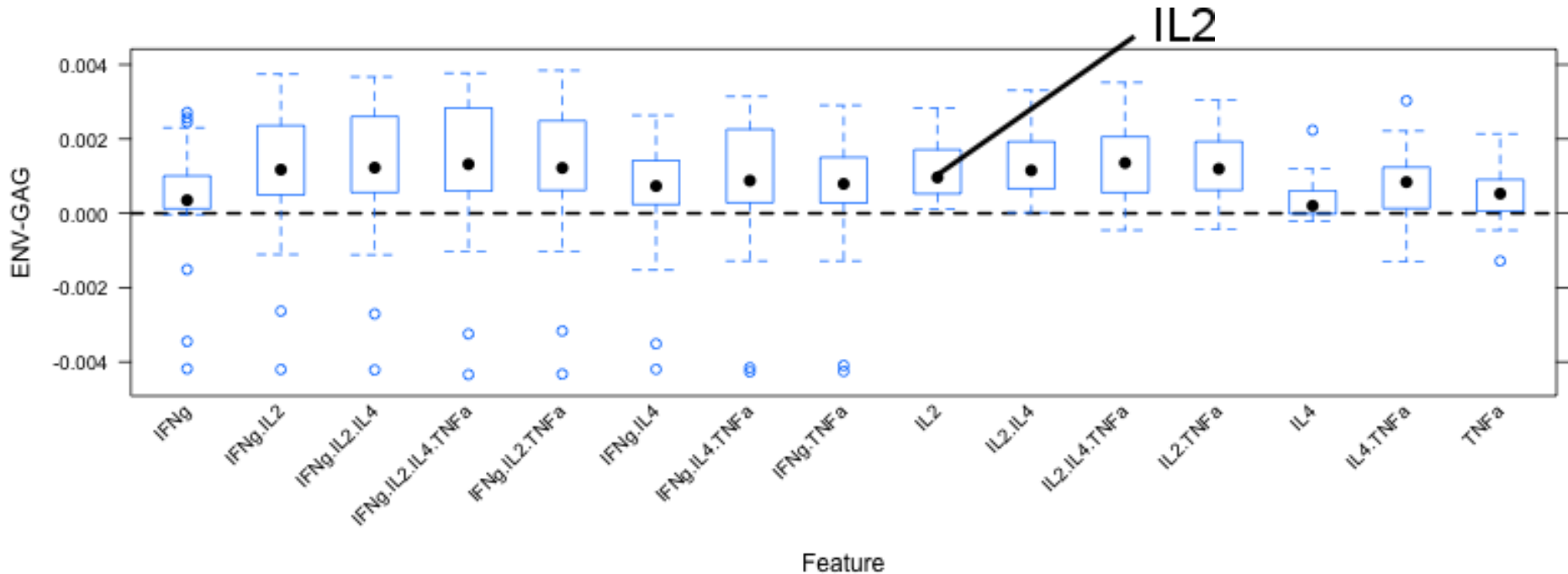
- Major peak modeled using robust mean and standard deviation.
- Outlier detection in the +ive direction used to identify positive cells.



RangeGate on Cytokine channels of CD4+



Env/Gag classification



- Use proportion of Cytokine+ cells as features
 - 4 from cd4+
 - 4 from cd8+
- Use paired data.
 - -If one sample of a pair is ENV, the other is necessarily GAG.
 - ENV is systematically higher than GAG for each sample pair in the training data.

Responder/ non-responder calls

- For **each patient**, does a sample respond to the stimulation
- The usual approach is to use **Fisher's exact test** on the count data.
- We take a Bayesian approach...
- Fit a standard Beta-Binomial model to raw counts from each stimulation/control pair.
- Estimate the posterior probability that the proportion of stimulated cells is greater than the control.

Beta-Binomial Model

Negative Positive

Stimulated	C_i^{s-}	C_i^{s+}	C_i^{st}	$C_i^{c+} \sim \mathbf{Bin}(C_i^{ct}, p_i^c)$
Unstimulated	C_i^{c-}	C_i^{c+}		

Prior: $p_i^c \sim \mathbf{Be}(\alpha_s, \beta_s)$

κ shrinkage factor

α_s, β_s are estimated from the data

Posterior

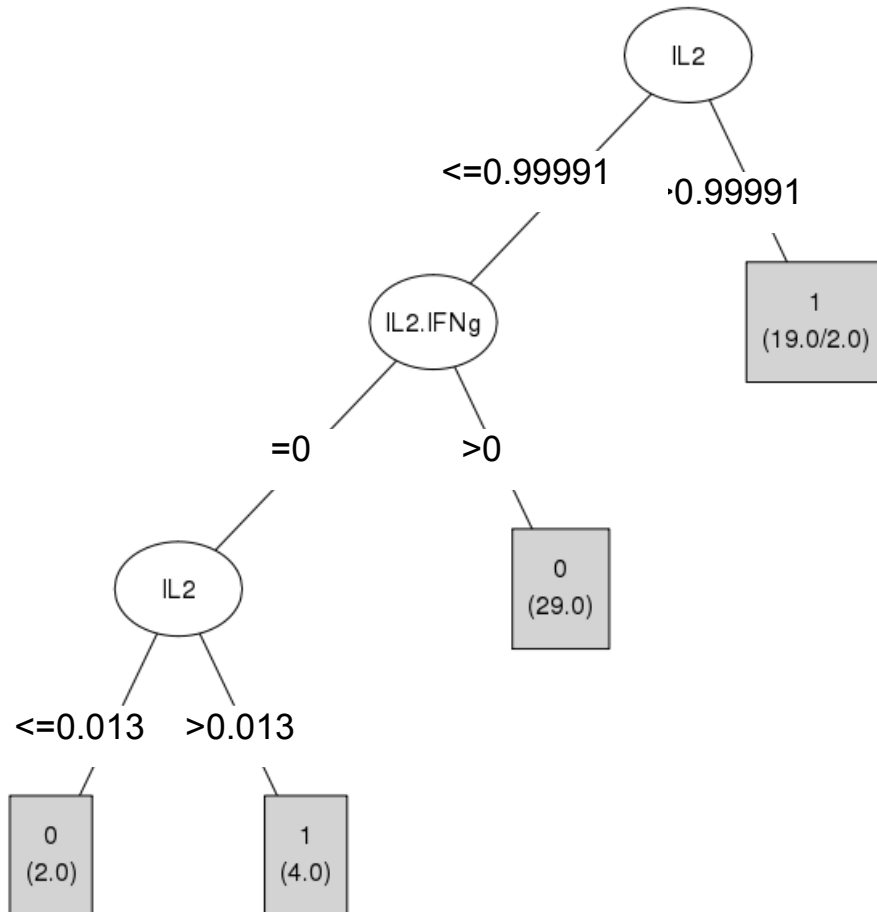
$$f(p_i^s | C_i^{s+}, C_i^{s-}) = \mathbf{Be}(p_i^s | C_i^{s+} + \alpha_s \kappa, C_i^{s-} + \beta_s \kappa)$$

$$f(p_i^c | C_i^{c+}, C_i^{c-}) = \mathbf{Be}(p_i^c | C_i^{c+} + \alpha_s \kappa, C_i^{c-} + \beta_s \kappa)$$

Finally, estimate: $P(p_i^s > p_i^c)$ via Monte Carlo

Response Prediction

- Calibrate the posterior probabilities using the training data.
- Decision tree to choose cutoff and features.



Features used for classification were IL2, and IFNg|IL2

2 non-responders misclassified as responders on training data.