Analyzing Flow-Cytometry Count Data with Regression Mixtures

Amit Meir
University of Washington

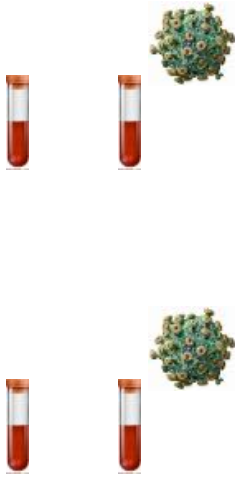
Joint work with

Raphael Gottardo and **Greg Finak**
Fred Hutchinson Cancer Research Center

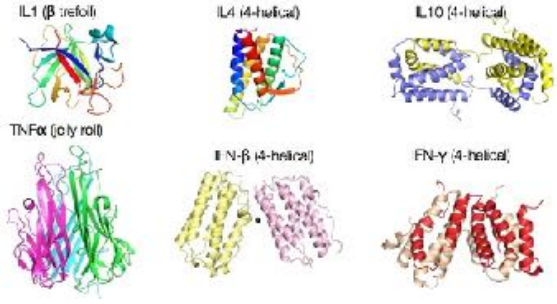
June 24, 2017

The RV144 HIV Vaccine Trial

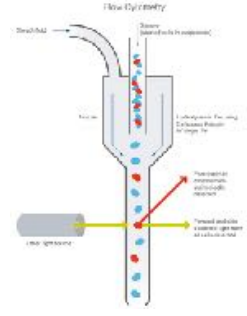
1. Stimulation



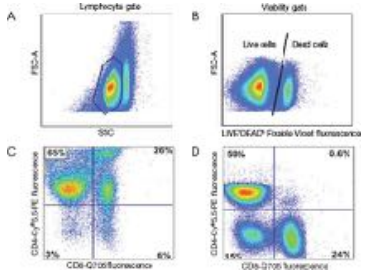
2. Response



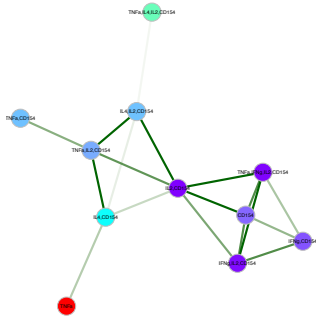
3. Flow-Cytometry



4. Gating



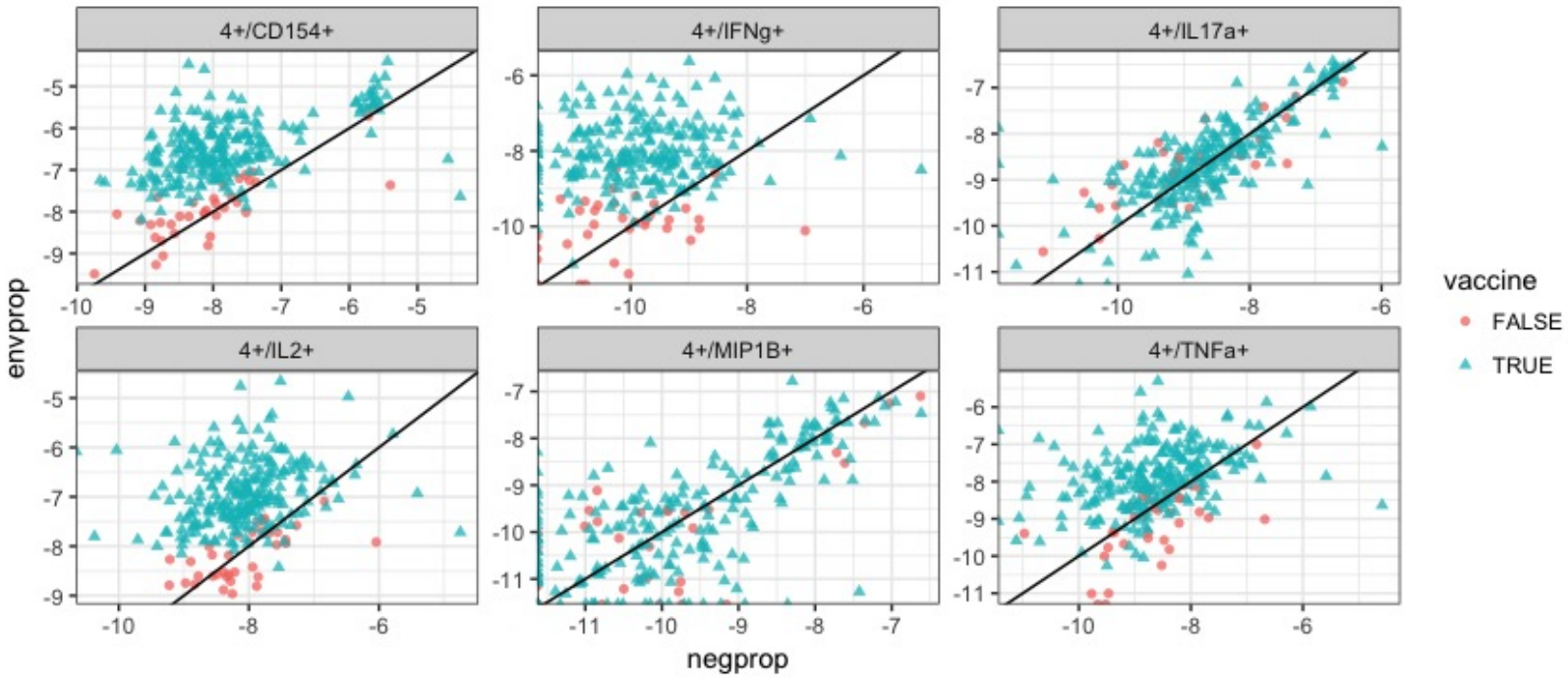
5. Analysis



The RV144 HIV Vaccine Trial

PTID	Subset	stim	count	parentcount
P1003	CD154	stim	38	23524
P1003	CD154	nonstim	31	28099
P1003	CD154,IL17a	stim	23	23524
P1003	CD154,IL17a	nonstim	30	28099
P1003	IFNg	stim	1	23524
P1003	IFNg	nonstim	0	28099
P1003	IFNg,CD154	stim	1	23524
P1003	IFNg,CD154	nonstim	0	28099
P1003	IFNg,IL2	stim	2	23524
P1003	IFNg,IL2	nonstim	0	28099
P1003	IFNg,IL2,CD154	stim	0	23524
P1003	IFNg,IL2,CD154	nonstim	0	28099
P1003	IFNg,IL4,IL2,CD154	stim	0	23524
P1003	IFNg,IL4,IL2,CD154	nonstim	0	28099

Analysis Goals



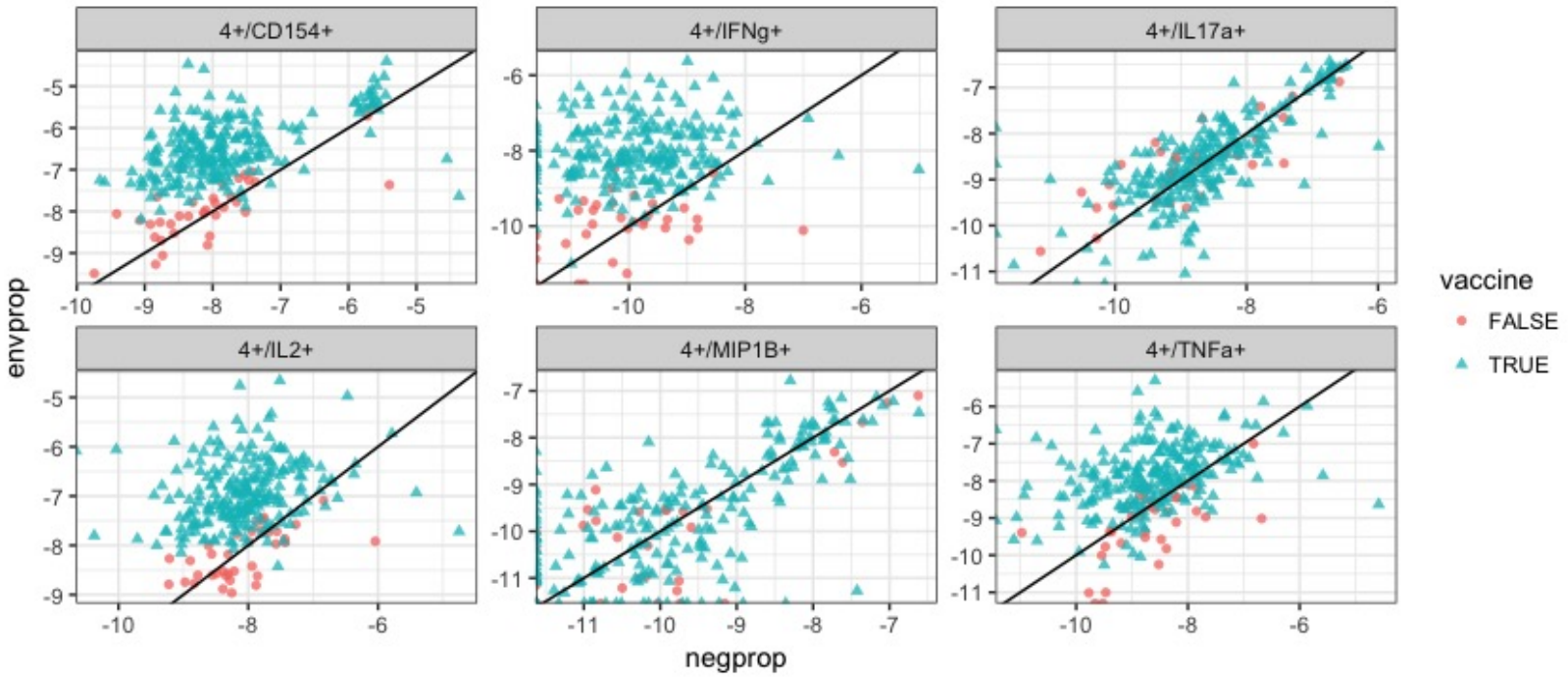
- Identify cell-subset that exhibit **vaccine specific response**
- Identify correlates for succesful **immunization**
- Infer **dependence** structures

Why a Regression Framework?

Current solutions are all based on comparing a single control sample to a stimulated sample.

- **Beyond baseline/stimulation**
  - Longitudinal data
  - Multiple stimulations per subject
- **Covariates**
  - Batch effects
  - Demographic/background information
- **Explicit Dependence Model**

Challenges



- **Overdispersion** (compared to Binomial)
- Subject specific **baseline response**
- Different **response** patterns
- **Dependence** across sub-samples AND cell-subsets
- **Dimensionality:** 100+ cell-subsets.

A Regression Model

Indexing: **i**-subject, **t**- subsample, **j**- subset.

- Over-dispersion $\Rightarrow$ **Beta-Binomial** model:

$$\text{logit}(\mu_{ijt}) = X_{ijt}\beta + T_{ijt}z_{ij} + \nu_{ij}$$

$$y_{ijt} \sim \text{Beta-Binom}(N_{it}, \mu_{ijt}, M_j)$$

- X- Covariates, β- Regression Coefficients, T- Treatment Effects.
- Baseline response $\Rightarrow \nu_i \sim N(0, \Sigma)$
- Differential response $\Rightarrow z_i \in \{0,1\}^J \sim \textit{Ising}(\theta)$.
- Estimation via a Stochastic-EM algorithm.

A Regression Model

Indexing: **i**-subject, **t**- subsample, **j**- subset.

- Over-dispersion $\Rightarrow$ **Beta-Binomial** model:

$$\text{logit}(\mu_{ijt}) = X_{ijt}\beta + T_{ijt}z_{ij} + \nu_{ij}$$

$$y_{ijt} \sim \text{Beta-Binom}(N_{it}, \mu_{ijt}, M_j)$$

- X- Covariates, β- Regression Coefficients, T- Treatment Effects.

- Baseline response $\Rightarrow \nu_i \sim N(0, \Sigma)$

- Differential response $\Rightarrow z_i \in \{0,1\}^J \sim \textit{Ising}(\theta)$.

- Estimation via a Stochastic-EM algorithm.

The RV144 HIV Vaccine Trial

- **262 Subjects**
  - 226 Cases
  - 36 Controls
- **2 Types of stimulus**
  - HIV protein
  - Negative control
- **Demographic Information**
  - Age
  - Gender
- **23 CD4 Cell-Subsets.**

The Plan

1. Fit the model based on **count data**:

   $$\text{logit}(\mu_{ijt}) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{gender}_i + z_{ij}\tau_j \text{stimulation}_{ijt} + \nu_{ij}$$

   Outputs:
   - Regression coefficient estimates
   - Posterior response probabilities
   - Covariance for random effects
   - Estimated graphical model

2. Validate inferred quantities using **vaccination** data.

3. Formulate hypothesis and test using **infection** data.

The Plan

1. Fit the model based on **count data**:

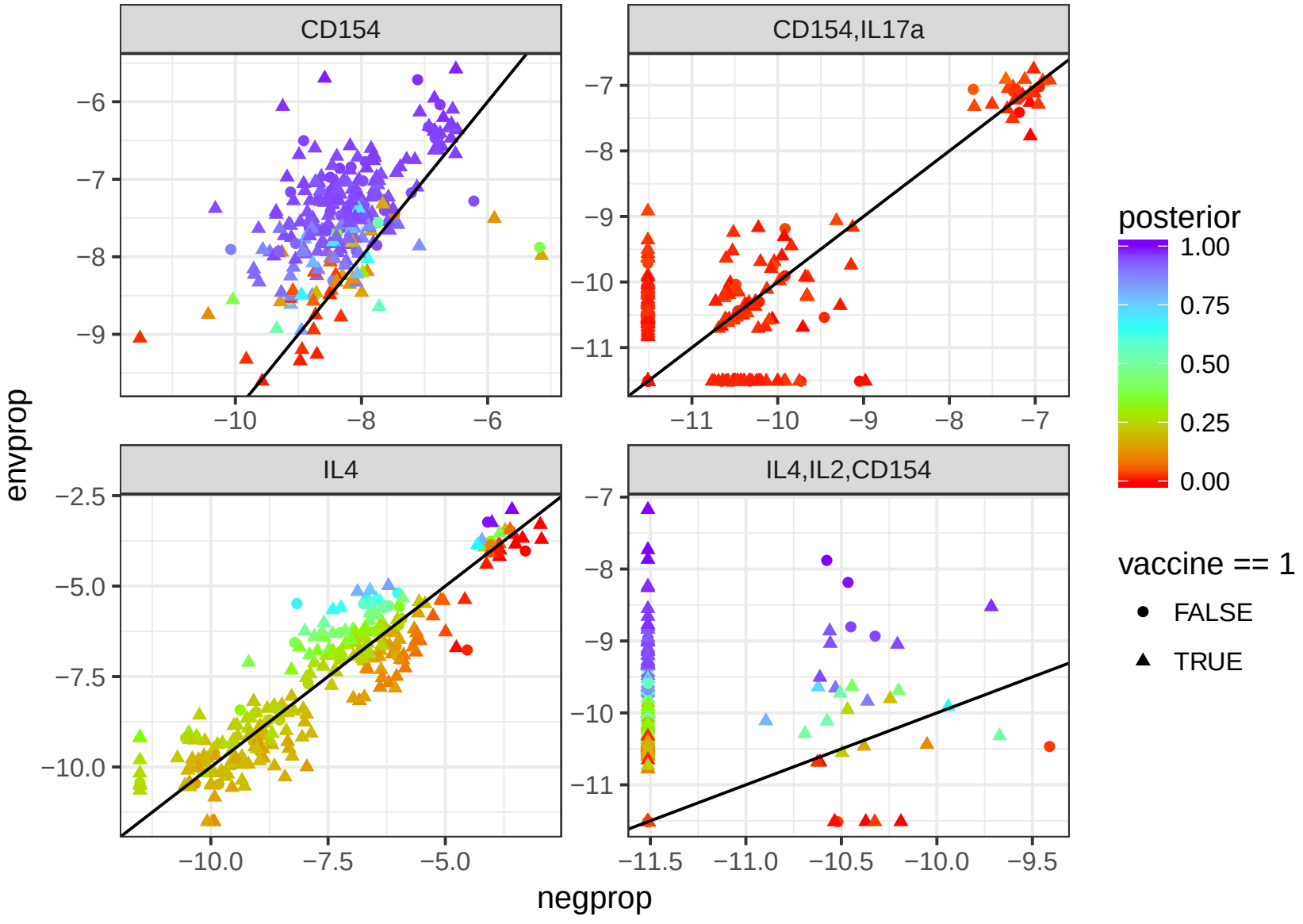
   $$\text{logit}(\mu_{ijt}) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{gender}_i + z_{ij}\tau_j \text{stimulation}_{ijt} + \nu_{ij}$$

   Outputs:
   - Regression coefficient estimates
   - Posterior response probabilities
   - Covariance for random effects
   - Estimated graphical model

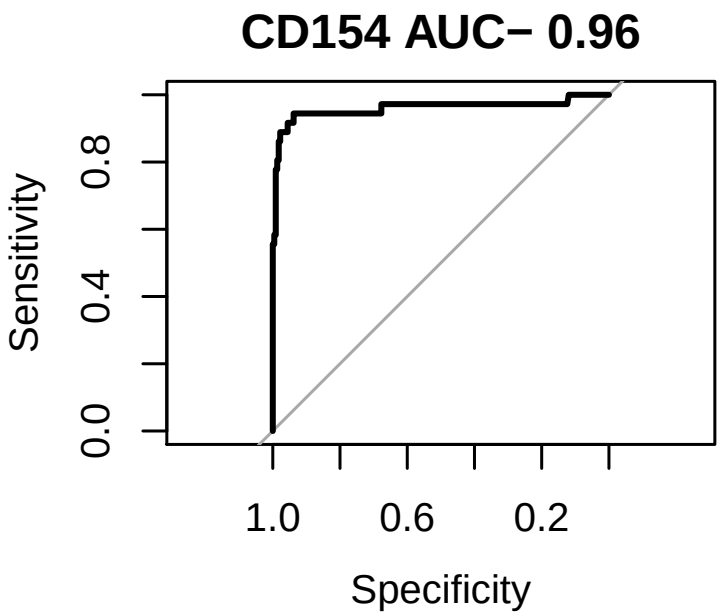
2. Validate inferred quantities using **vaccination** data.

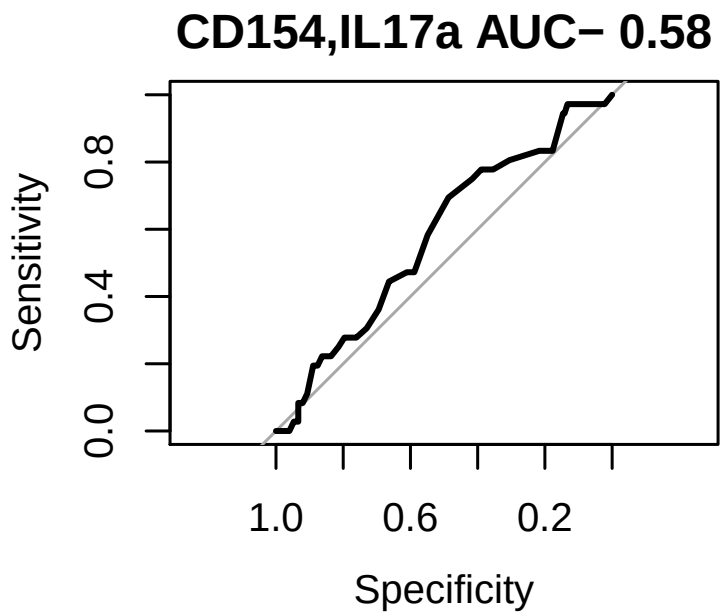
3. Formulate hypothesis and test using **infection** data.

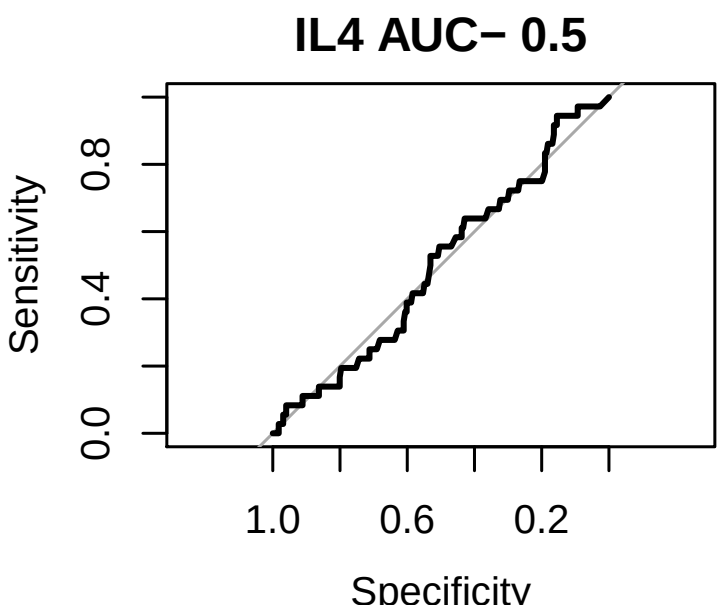
RV144 - Booleans Dataset

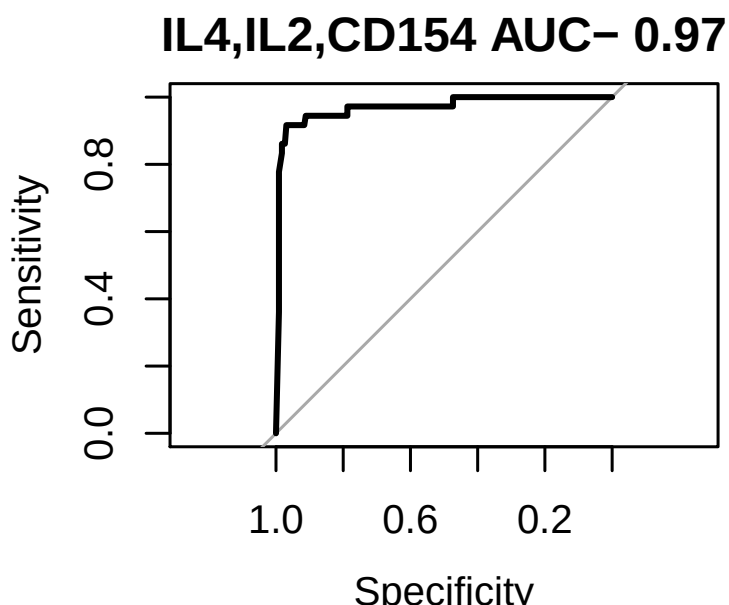


RV144 - Booleans Dataset

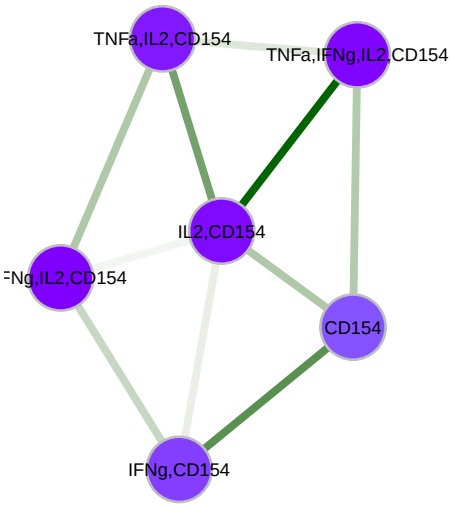


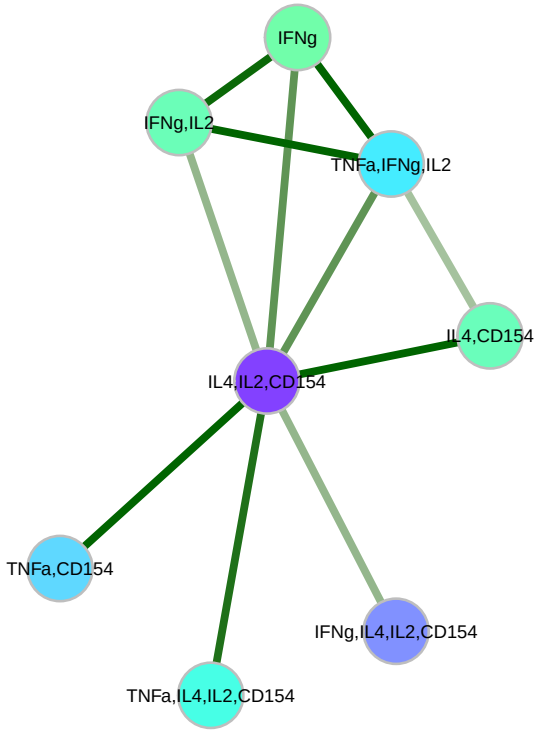


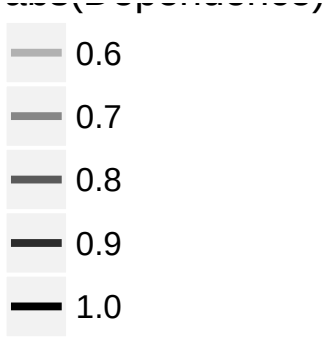


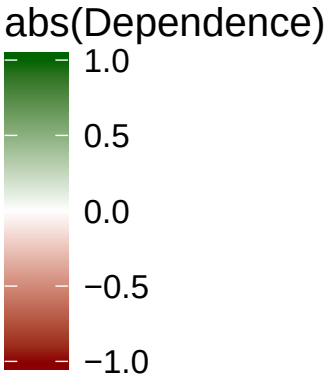


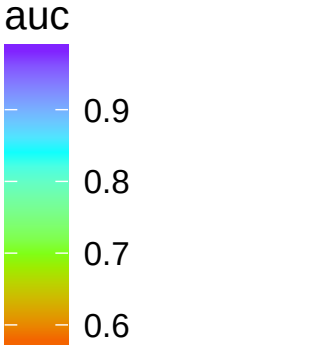
An Informative Graphical Model



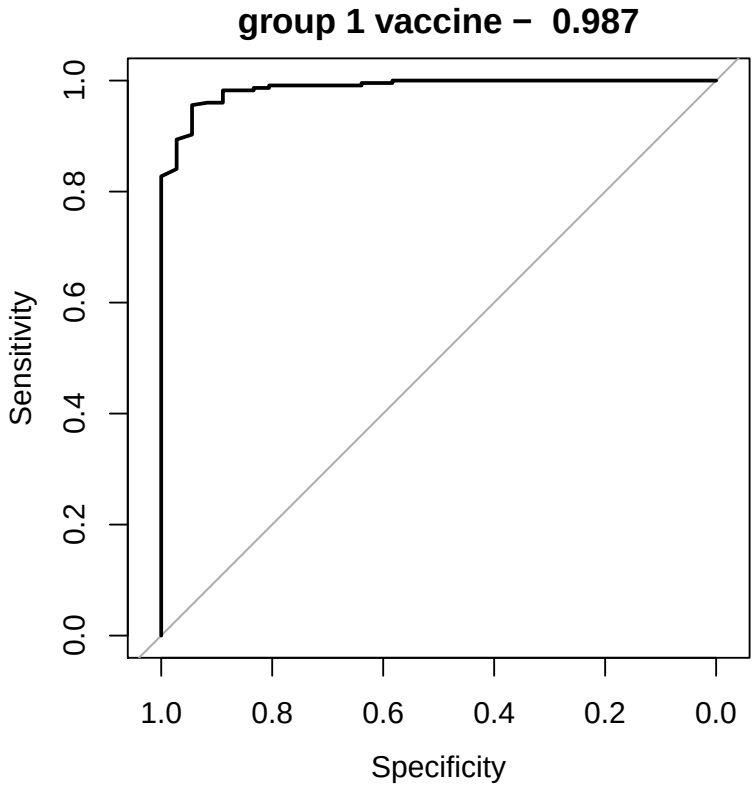


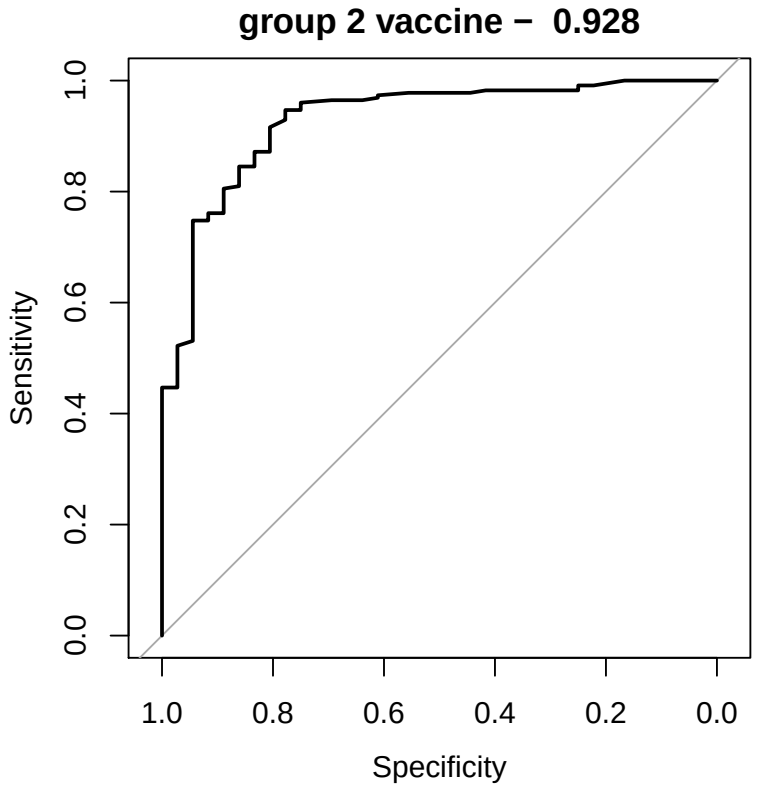






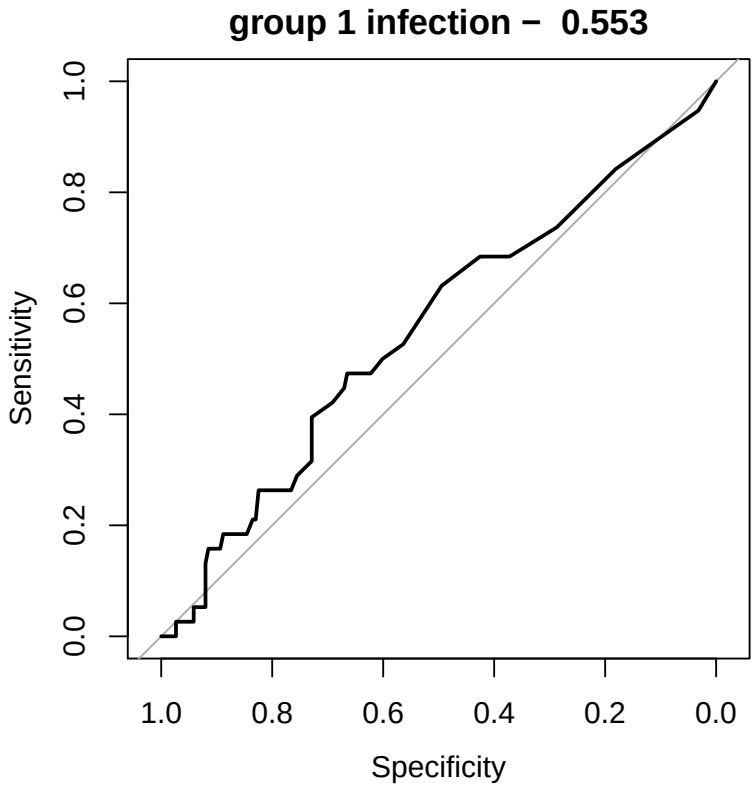
ROC for Vaccination/Placebo

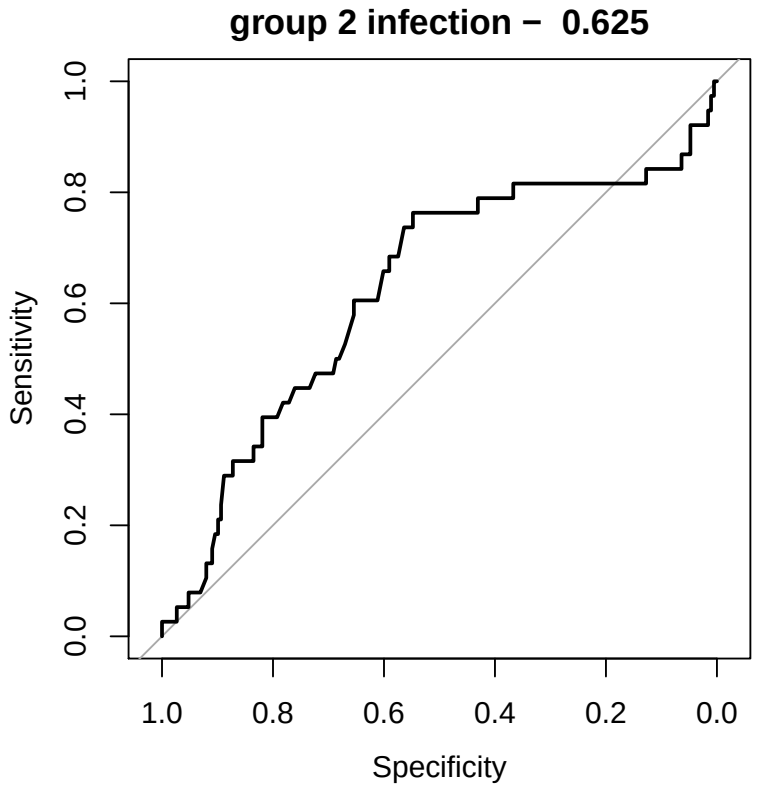




$$\text{ROC}(\text{vaccine} \sim s_j), \qquad s_{ij} = \frac{1}{|C_j|} \sum_{i \in C_j} \text{post}_{ij}$$

ROC for Infection Status



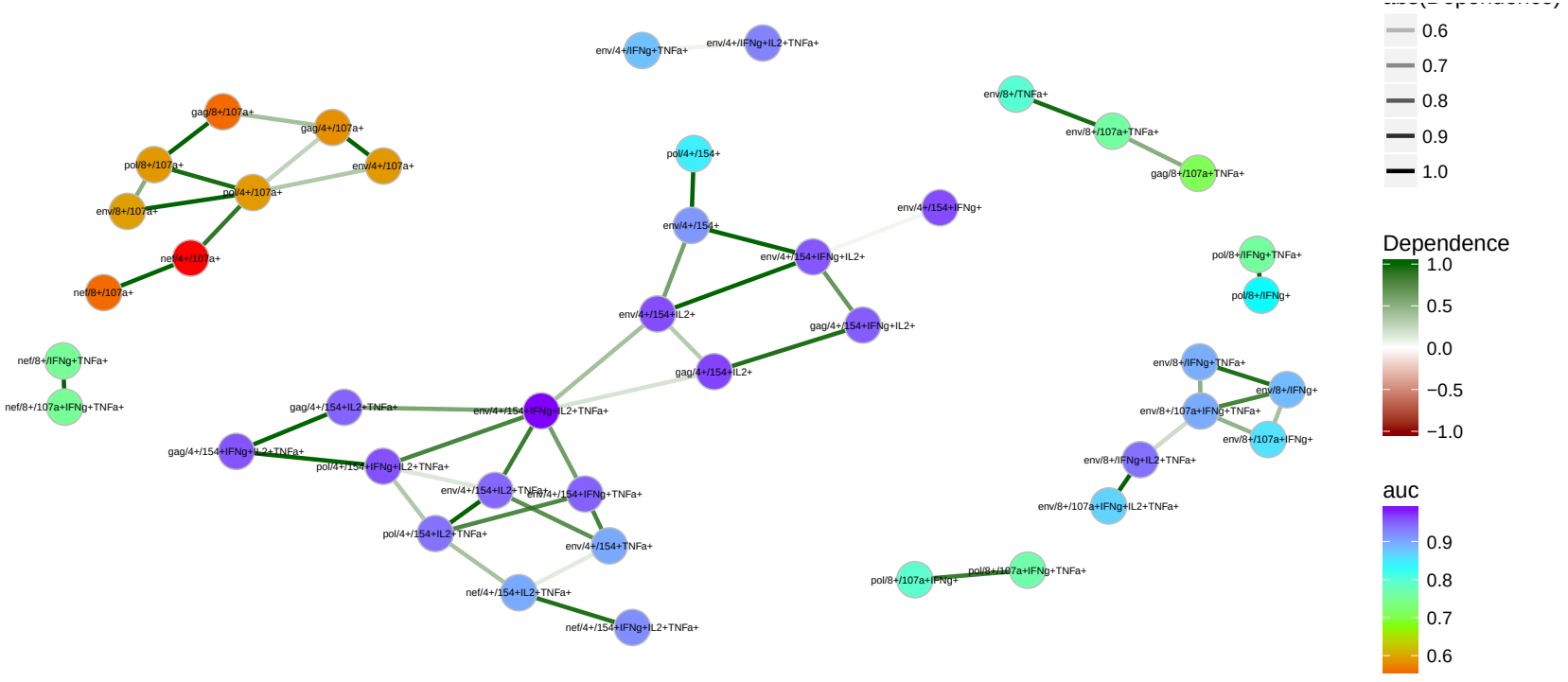


AUC of 0.625 $\Rightarrow$ p-value of 0.007.

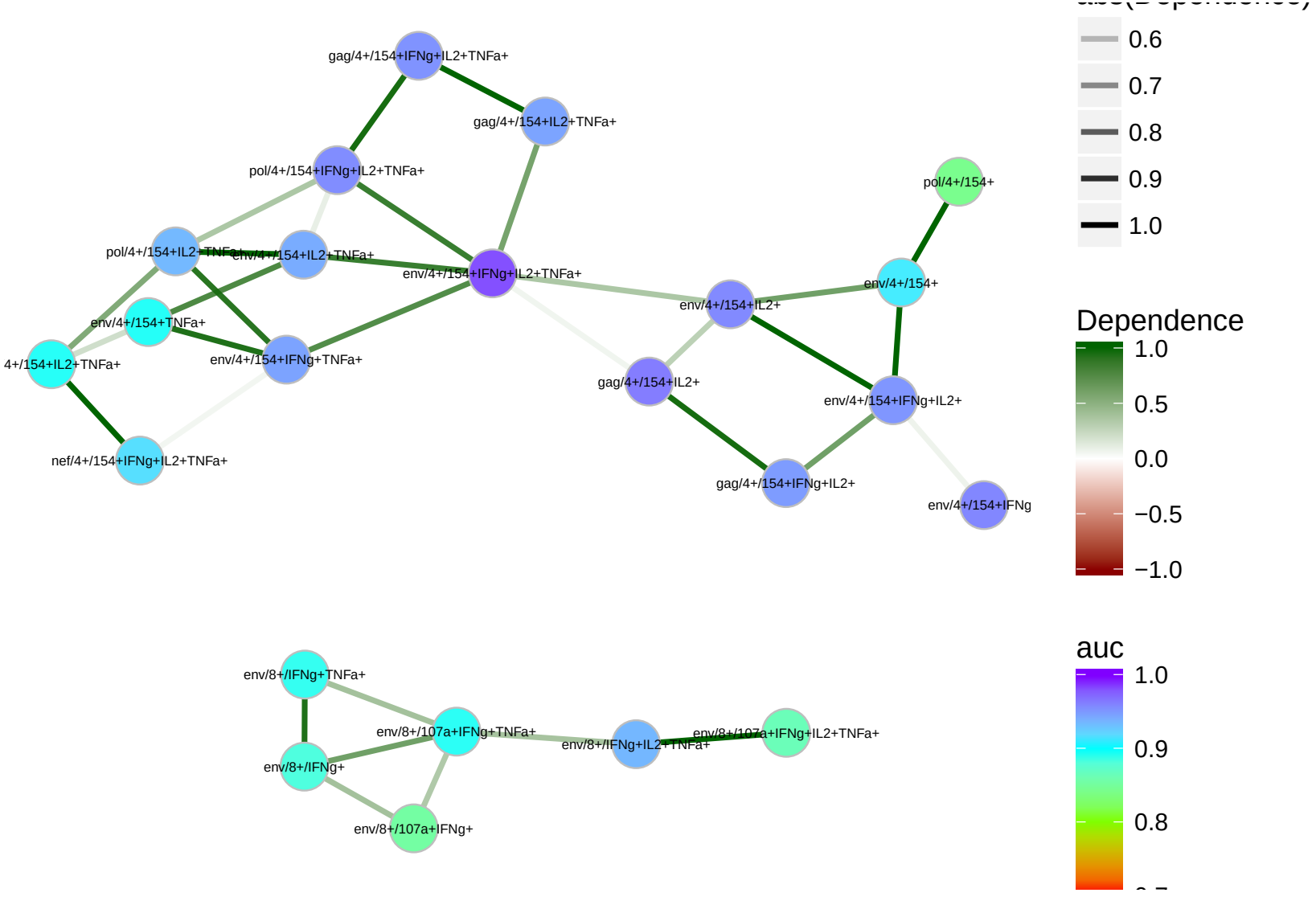
The HVTN 505 Vaccine Trial

- **238 Subjects**
  - 189 Cases
  - 49 Controls
- **5 Types of stimulus**
  - 4 types of HIV proteins (ENV, GAG, POL, NEF).
  - Negative control.
  - Multiple samples per stimulation.
- **52 Cell Subsets**
  - 25 CD4 cells.
  - 27 CD8 cells.

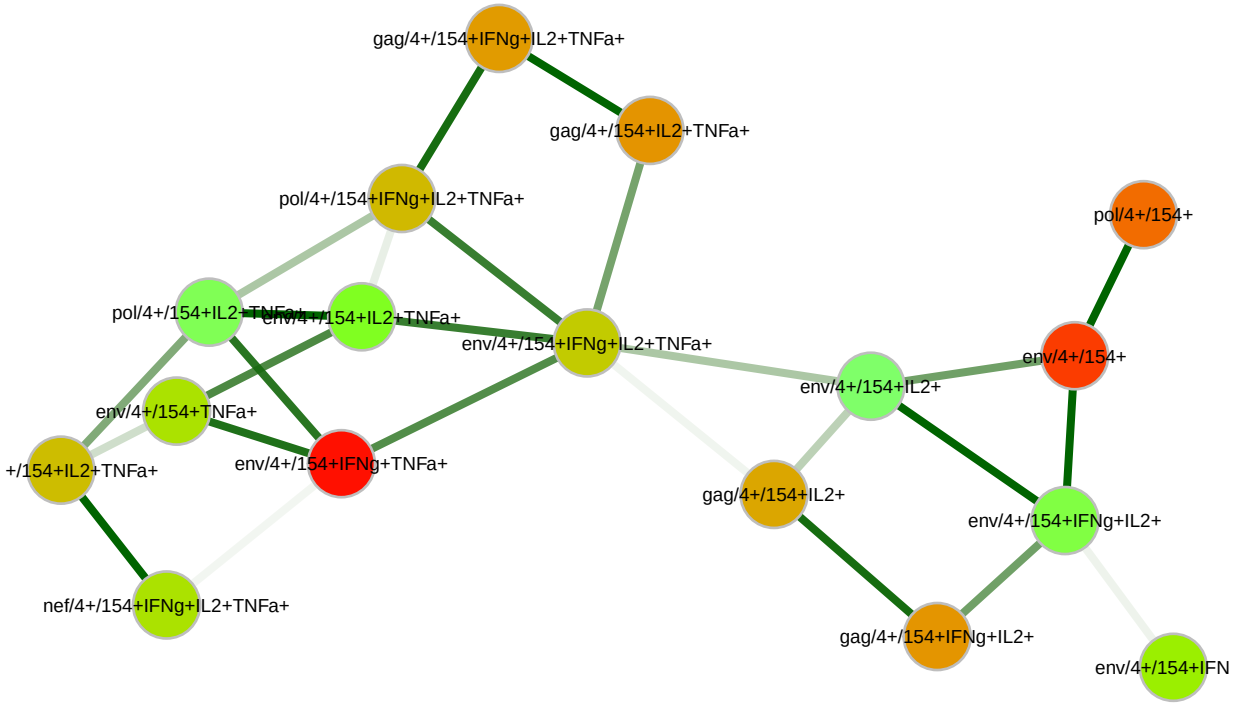
Inferred Graph for HVTN505

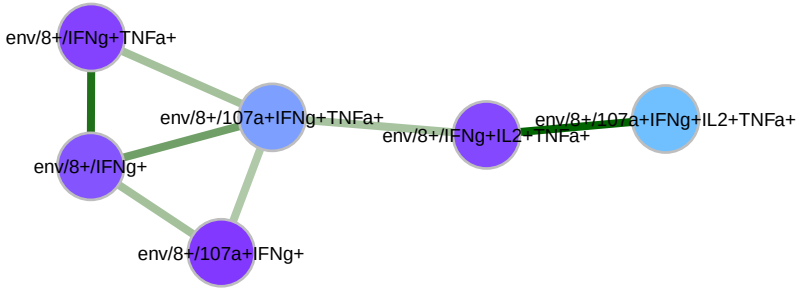


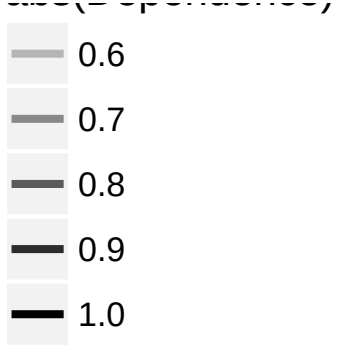
Color Coded by AUCs for Vaccination/Placebo

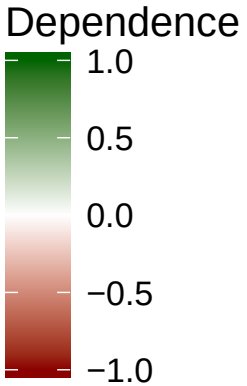


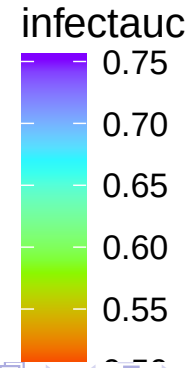
Color Coded by AUCs for Infection Status



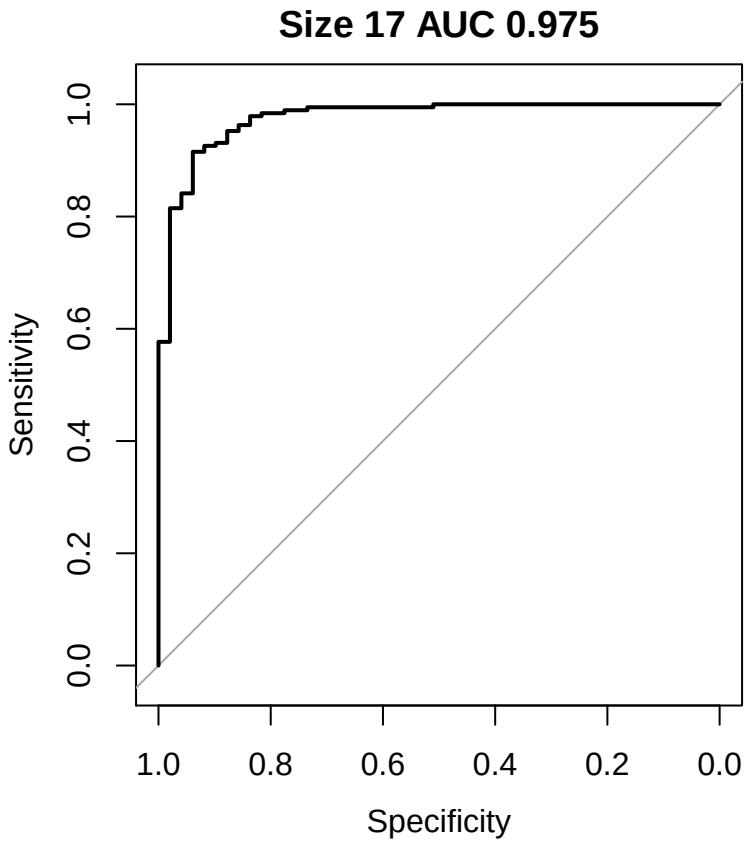


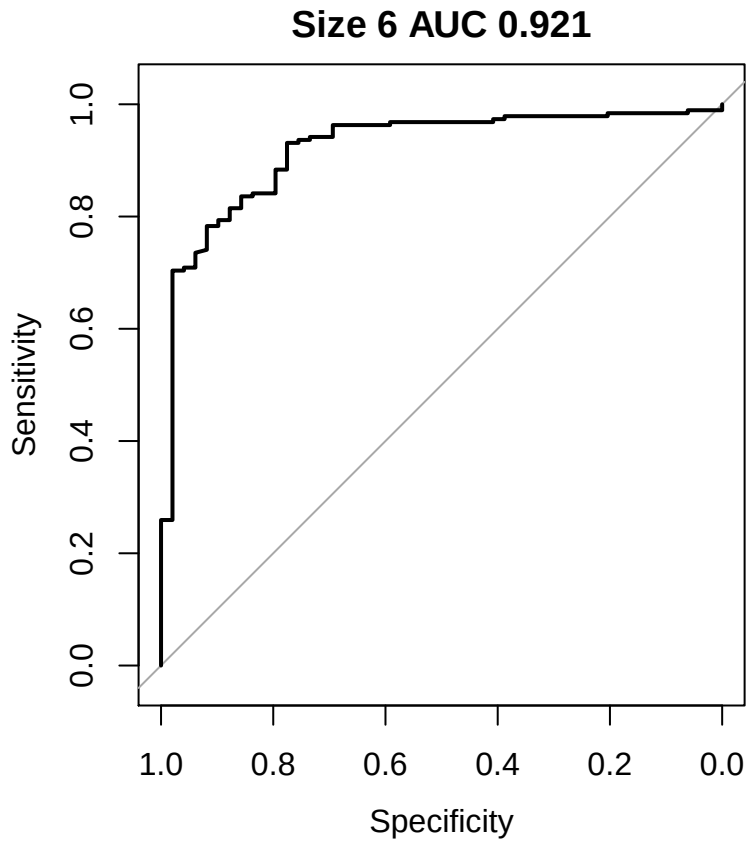




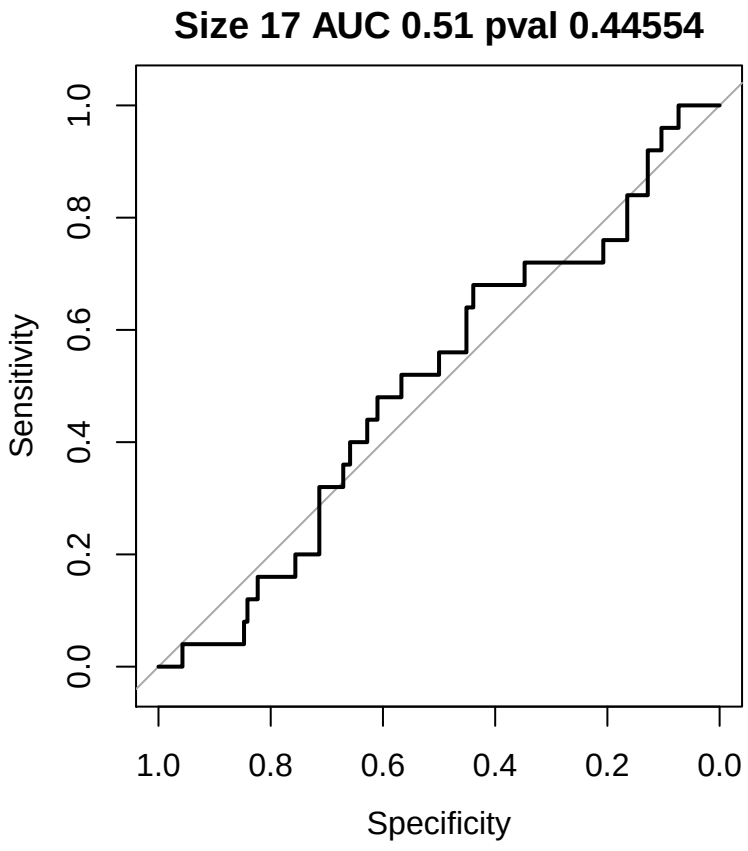


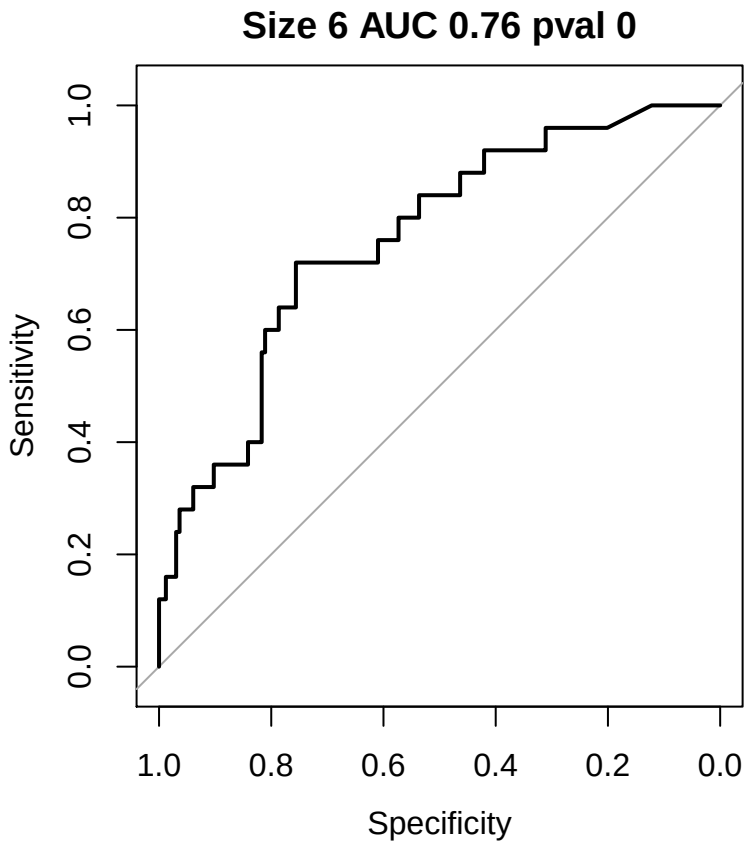
ROC for Vaccination/Placebo





ROC for Infection Status





AUC of 0.76 $\Rightarrow$ p-value of $\approx 10^{-8}$.

Conclusion

- We developed a regression model which allows for the analysis of complex cell-count datasets.
  - Multiple time-points/observations per subject.
  - Batch effects
  - Demographic Information
- We model the dependence structure explicitly via a sparse graphical model.
  - Identified subsets predictive of vaccination **or** immunization.
- What else?
  - Longitudinal data
  - Enrichment Analysis
  - Aggregate measures of response

Thank you!

Questions?

AmitMeir@uw.edu

Analysis Goals

- **Problem:** We are interested in identifying response in Subsets X Protein pairs.
- **Solution:** Treat each combination of Subset X Protein as a cell-subset.
  - Overall 120 subsets with non-negligible counts.
- Dependence structures should (and do!) sort themselves out.

RV144 - Booleans Dataset