Analyzing Flow-Cytometry Count Data with Regression Mixtures

Amit Meir
University of Washington

Joint work with

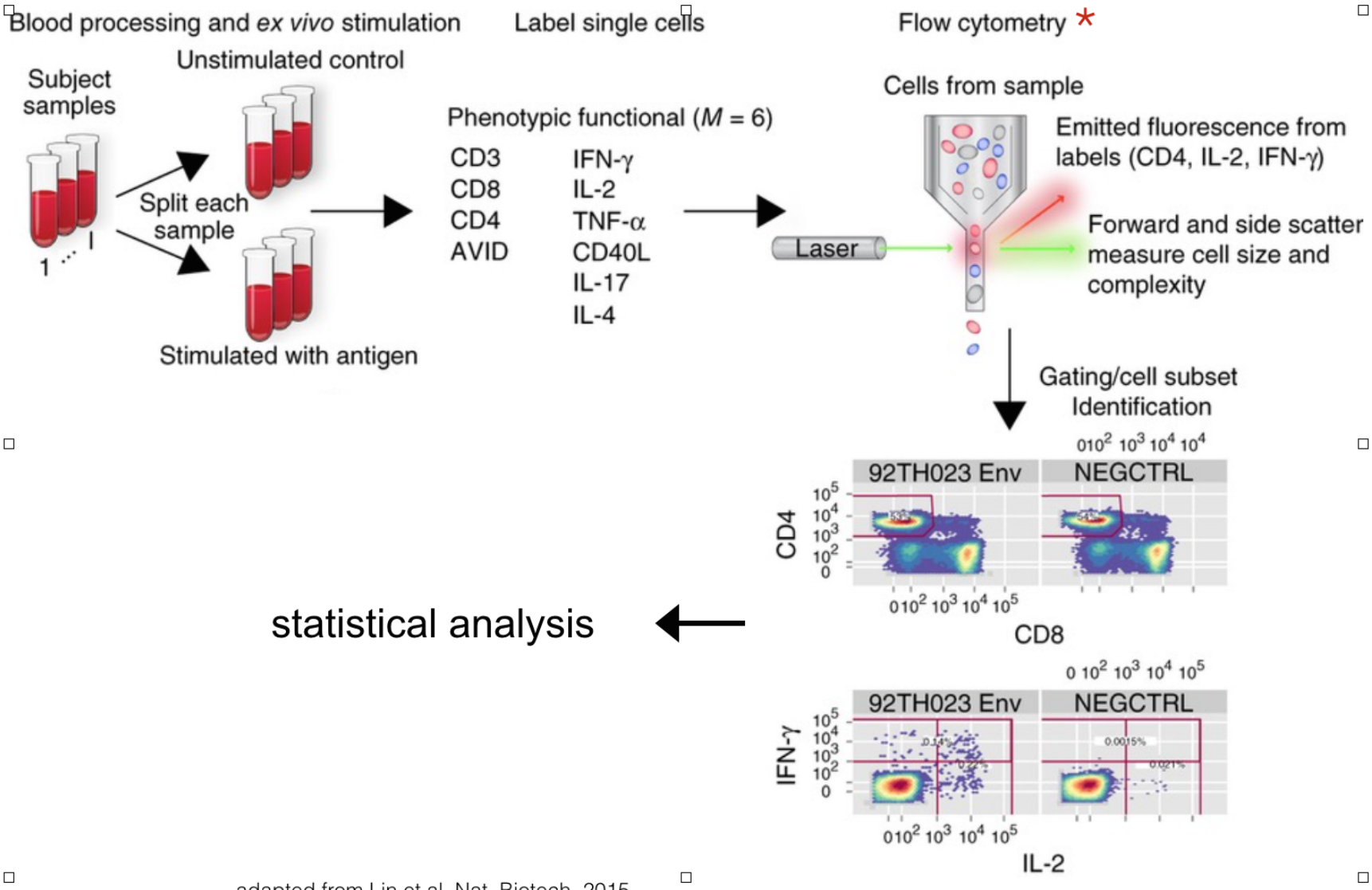
Raphael Gottardo and **Greg Finak**
Hutchinson Cancer Research Center

May 10, 2017

Outline

1. **Flow-Cytometry - a Refresher**
   - Flow Cytometry?????
   - A model in need of a better name than flowReMix.
2. **RV144**
   - Inferred Graphical Model
   - Cumulative Response Measures.
3. **HVTN505**:
   - Graphical Model.
   - Correlates for Infection Status (Clinical Outcome).
   - Breadth/Polyfunctionality Anaysis.
4. **CHMI Study**
   - Longitudal Data.
   - Enrichment Analysis.

The RV144 HIV Vaccine Trial

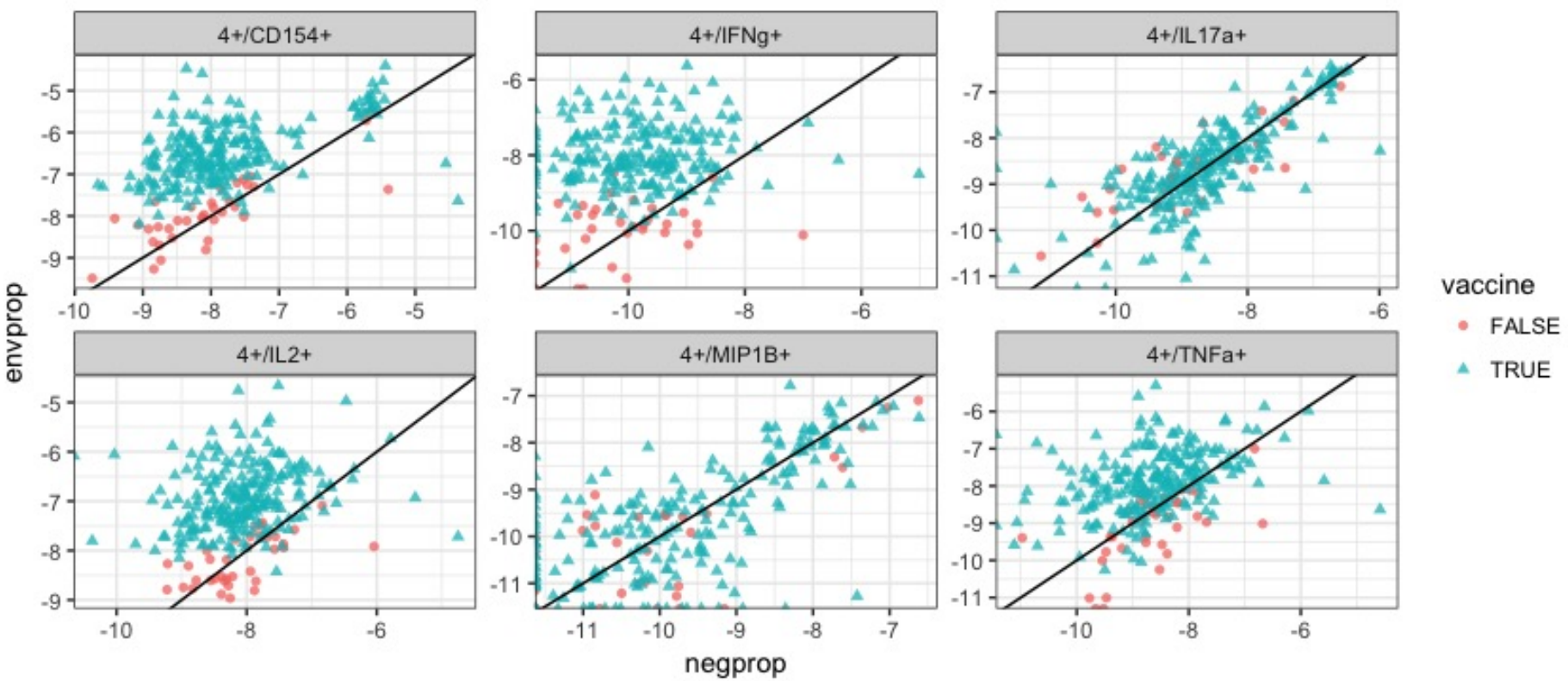


adapted from Lin et al, Nat. Biotech. 2015

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

Marginal Counts for RV144



The Beta-Binomial Distribution

The Beta-Binomial distribution is a type of over dispersed Binomial distribution.

$$p \sim Beta(\mu M, (1-\mu)M)$$

$$y \sim Bin(N, p)$$

$$E(\bar{y}) = \mu$$

$$Var(\bar{y}) = \frac{\mu(1-\mu)}{N} + \frac{\mu(1-\mu)}{M+1}$$

The Beta-Binomial Distribution

The Beta-Binomial distribution is a type of over dispersed Binomial distribution.

$$p \sim Beta(\mu M, (1-\mu)M)$$

$$y \sim Bin(N, p)$$

$$E(\bar{y}) = \mu$$

$$Var(\bar{y}) = \frac{\mu(1-\mu)}{N} + \frac{\mu(1-\mu)}{M+1}$$

A Random Intercept Model

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

$$\nu_i \sim N(0, \Sigma),$$

$$\text{logit}(\mu_{ijt}) = X_{ijt}\beta_j + \nu_{ij},$$

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

An Individual Response Model

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

We want to allow for individual subjects/cell-subsets to have differential response to stimulation.

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

$$\tau_{ij} = \begin{cases} \tau_j & \text{response in } \{i, j\} \\ 0 & \text{no response} \end{cases}.$$

A Markov Random Field Model

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

Denote cluster (Response) by a $z \in \{0,1\}^p$ vector with 1 indicating a responsive subset.

We assume an Ising model for the dependence structure between subsets:

$$P(z) \propto \sum_{j=1}^{p} z_j \theta_j + \sum_{u \neq v} z_u z_v \theta_{uv},$$

$$P(z_j = 1 | z_{-j}) = \theta_j + \sum_{u \neq j} z_u \theta_{uj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Markov Random Field Model

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

Denote cluster (Response) by a $z \in \{0,1\}^p$ vector with 1 indicating a responsive subset.

We assume an Ising model for the dependence structure between subsets:

$$P(z) \propto \sum_{j=1}^{p} z_j \theta_j + \sum_{u \neq v} z_u z_v \theta_{uv},$$

$$P(z_j = 1 | z_{-j}) = \theta_j + \sum_{u \neq j} z_u \theta_{uj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Markov Random Field Model

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

Denote cluster (Response) by a $z \in \{0,1\}^p$ vector with 1 indicating a responsive subset.

We assume an Ising model for the dependence structure between subsets:

$$P(z) \propto \sum_{j=1}^{p} z_j \theta_j + \sum_{u \neq v} z_u z_v \theta_{uv},$$

$$P(z_j = 1 | z_{-j}) = \theta_j + \sum_{u \neq j} z_u \theta_{uj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Hidden Markov Random Field Model

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

$$\nu_i \sim N(0, \Sigma),$$

$$z_i \sim \text{Ising}(\theta).$$

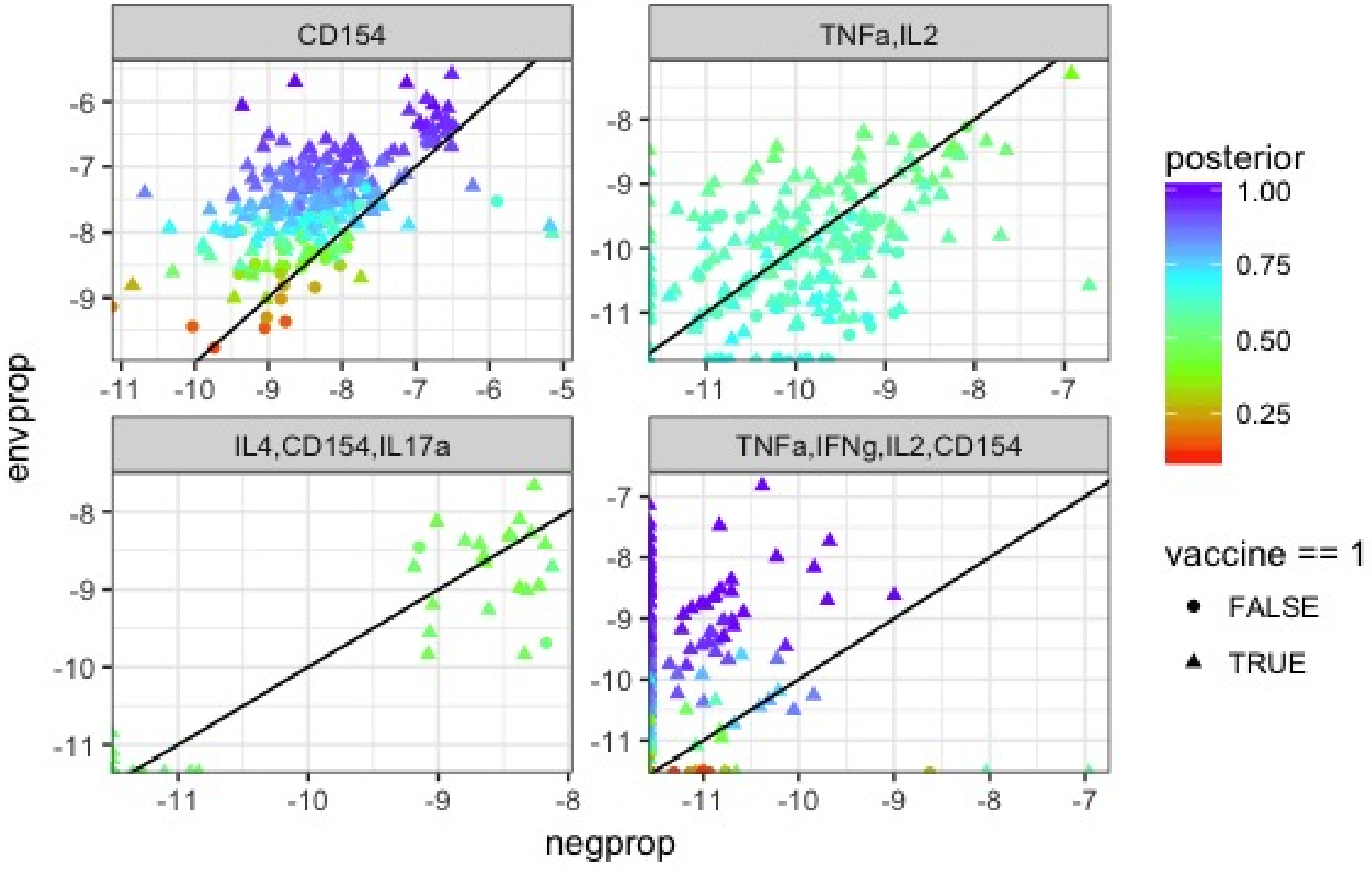
$$\text{logit}(\mu_{ijt}) = X_{ijt}\beta_j + T_{ijt}\tau_j(z_i) + \nu_{ij},$$

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

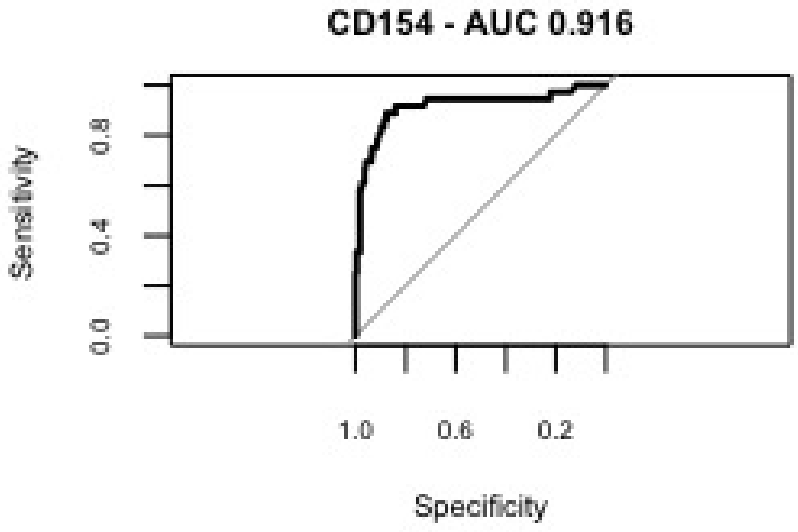
The RV144 HIV Vaccine Trial

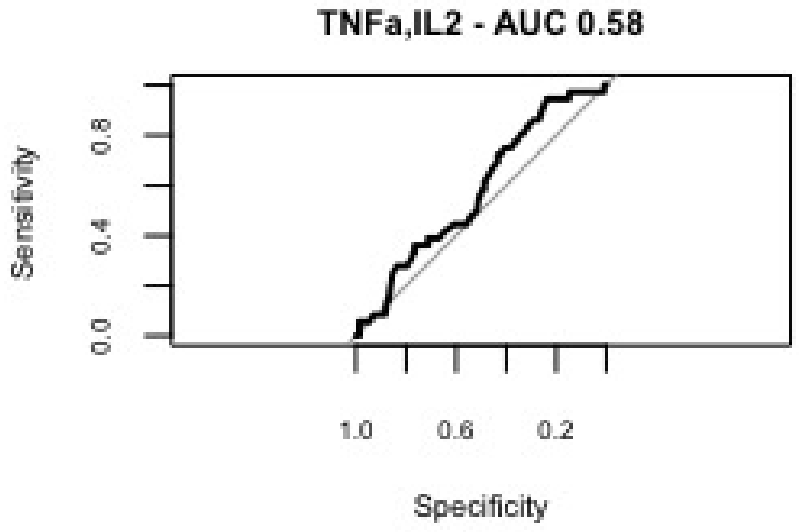
- **262 Subjects**
  - 226 Cases
  - 36 Controls
- **2 Types of stimulus**
  - HIV protein
  - Negative control
- **23 CD4 Cell-Subsets.**

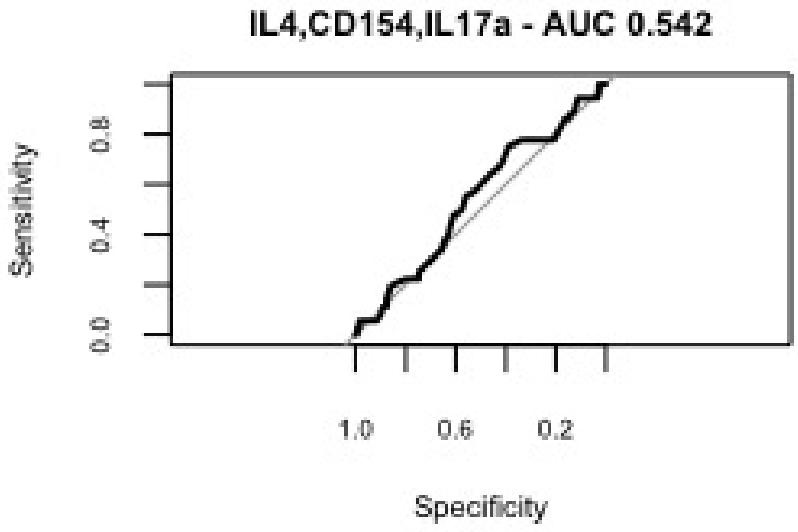
RV144 - Booleans Dataset

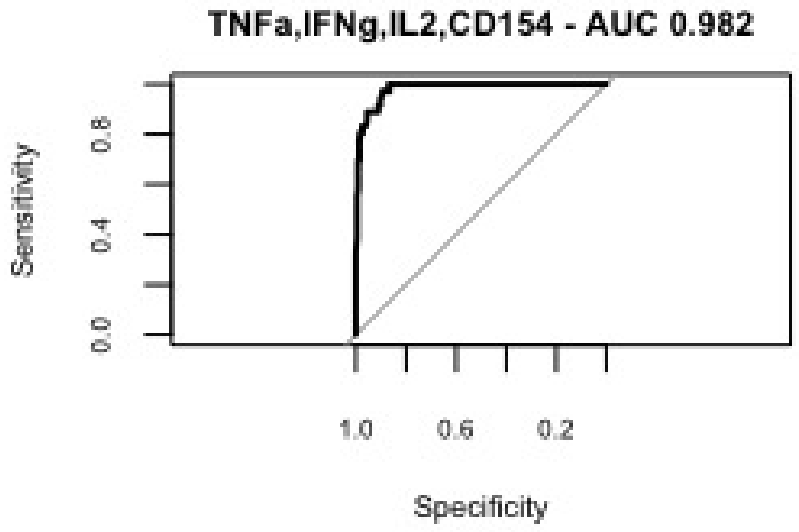


RV144 - Booleans Dataset









RV144 - Booleans Dataset

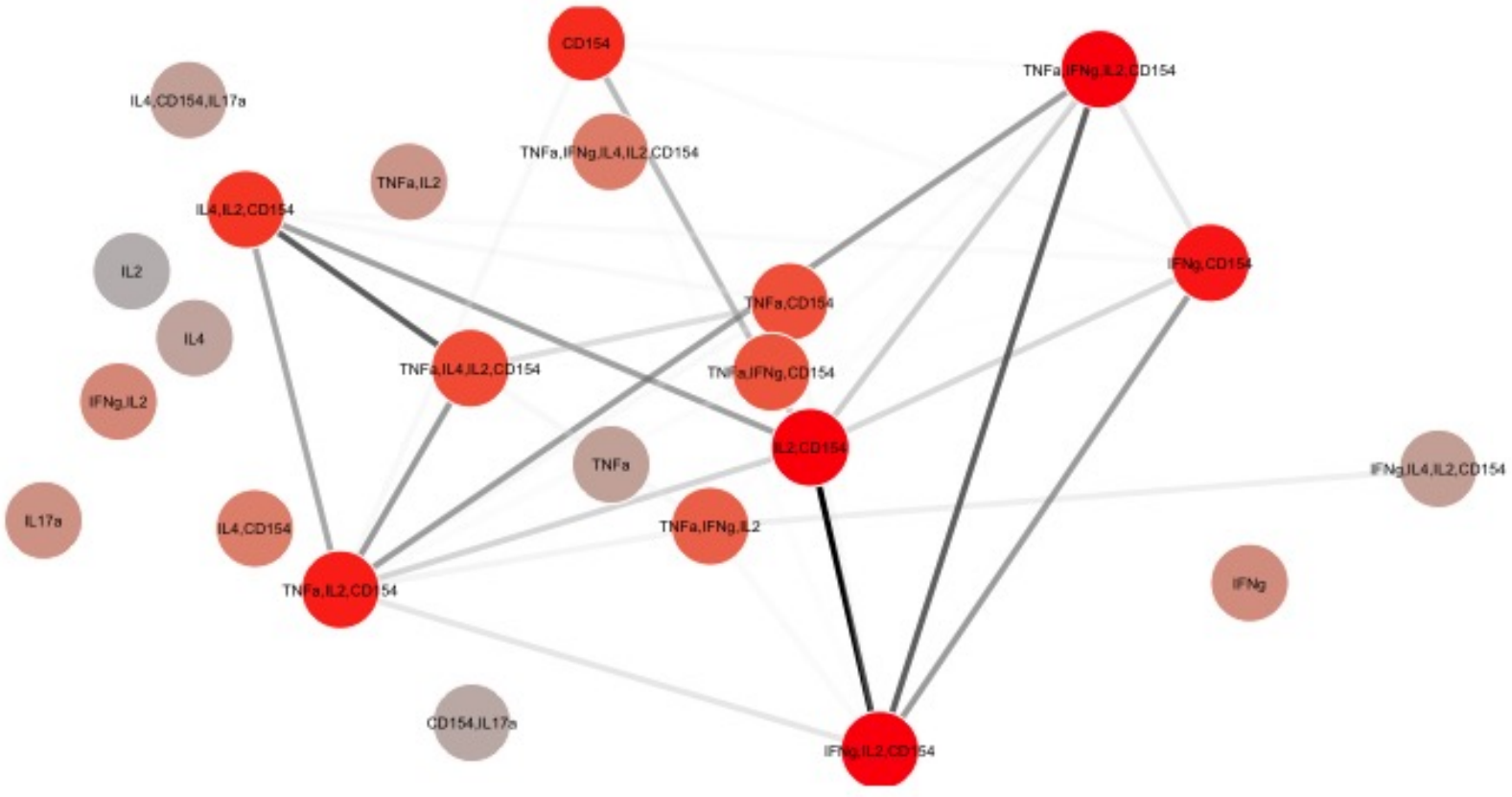


Figure: Estimated Ising Model - Red marks AUC

A Better Way to Estimate the Graphical Model?

The graph output by the procedure is an average of the graph estimated in several iterations.

- Not as sparse as we would like...
- How sure are we of existence of an edge?

Possible solution, stability selection:

- Draw samples from the posterior response distribution.
- Fit a Graphical Model.
- **Repeat**

Compute the proportion of estimated models in which an edge has been observed.

A Better Way to Estimate the Graphical Model?

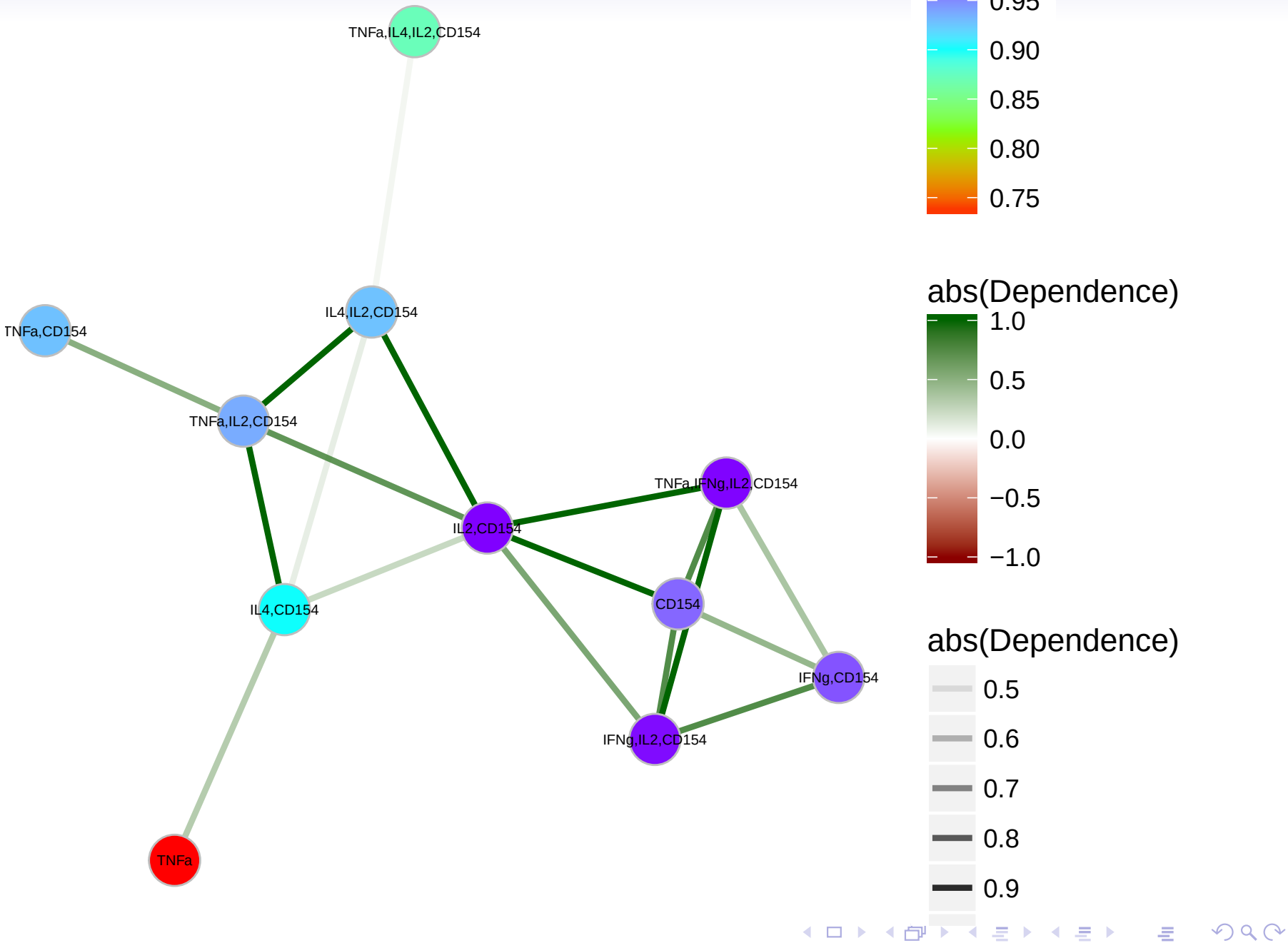
The graph output by the procedure is an average of the graph estimated in several iterations.

- Not as sparse as we would like...
- How sure are we of existence of an edge?

Possible solution, stability selection:

- Draw samples from the posterior response distribution.
- Fit a Graphical Model.
- **Repeat**

Compute the proportion of estimated models in which an edge has been observed.


TNFa,IL4,IL2,CD154
TNFa,CD154
IL4,IL2,CD154
TNFa,IL2,CD154
TNFa,IFNg,IL2,CD154
IL2,CD154
IL4,CD154
CD154
IFNg,CD154
IFNg,IL2,CD154
TNFa
0.95
0.90
0.85
0.80
0.75
abs(Dependence)
1.0
0.5
0.0
−0.5
−1.0
abs(Dependence)
0.5
0.6
0.7
0.8
0.9

Aggregating Subject Response

So far we have used posterior samples to:

- Identify responsive cell-subsets.
- Infer Dependence Structures.

How can we identify (or rank) responsive subjects?

- Responsive subject = 1 responsive subset? 2? 3?...
- How about stochastic ordering?

$$F \preceq G \quad \Leftrightarrow \quad G(x) \leq F(x) \ \ \forall x$$

We can compute a posterior CDF for # of responses for each subject!

Aggregating Subject Response

So far we have used posterior samples to:

- Identify responsive cell-subsets.
- Infer Dependence Structures.

How can we identify (or rank) responsive subjects?

- Responsive subject = 1 responsive subset? 2? 3?...
- How about stochastic ordering?

$$F \preceq G \quad \Leftrightarrow \quad G(x) \leq F(x) \ \ \forall x$$

We can compute a posterior CDF for # of responses for each subject!

Aggregating Subject Response

So far we have used posterior samples to:

- Identify responsive cell-subsets.
- Infer Dependence Structures.

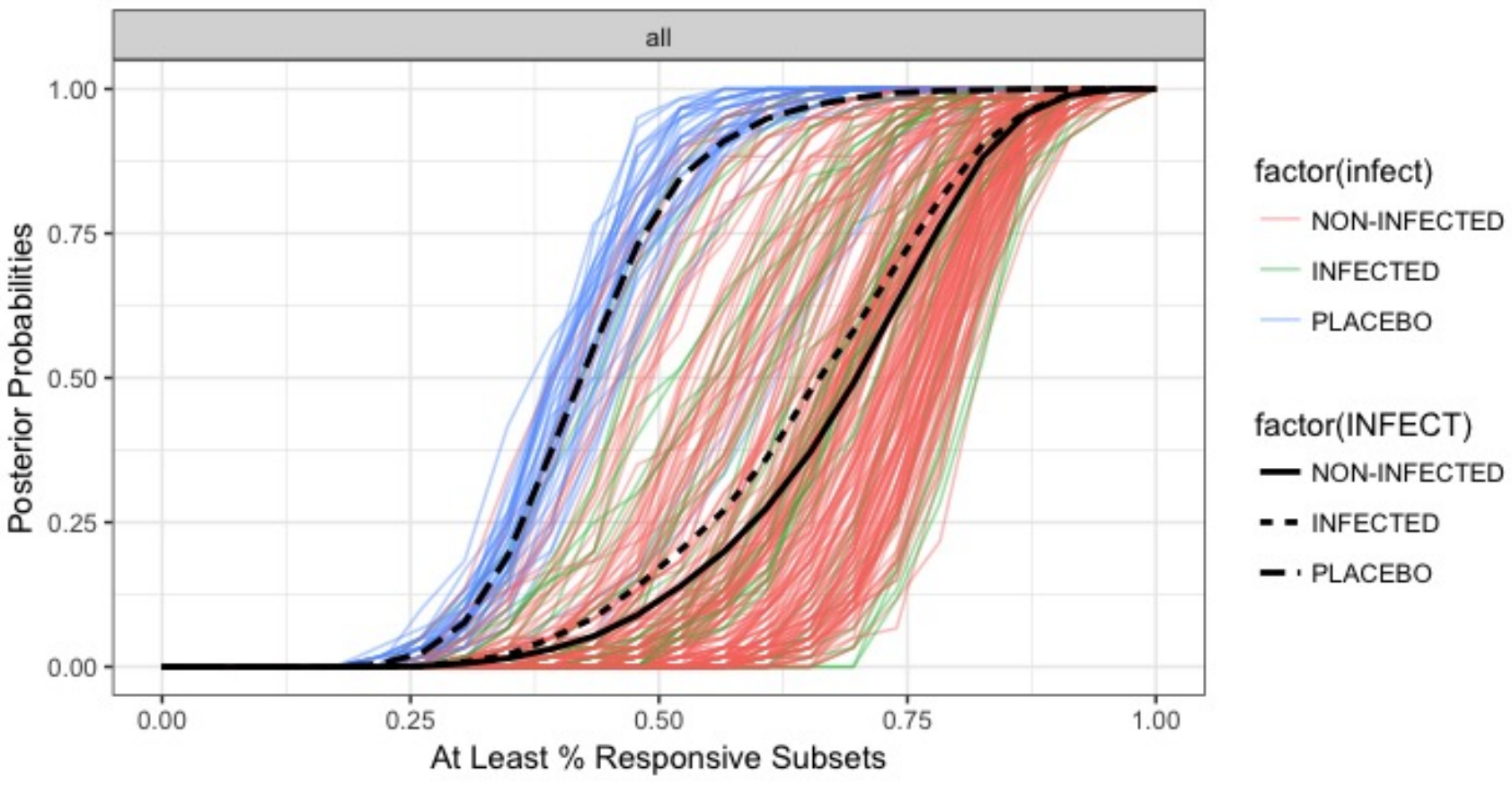
How can we identify (or rank) responsive subjects?

- Responsive subject = 1 responsive subset? 2? 3?...
- How about stochastic ordering?

$$F \preceq G \quad \Leftrightarrow \quad G(x) \leq F(x) \ \ \forall x$$

We can compute a posterior CDF for # of responses for each subject!

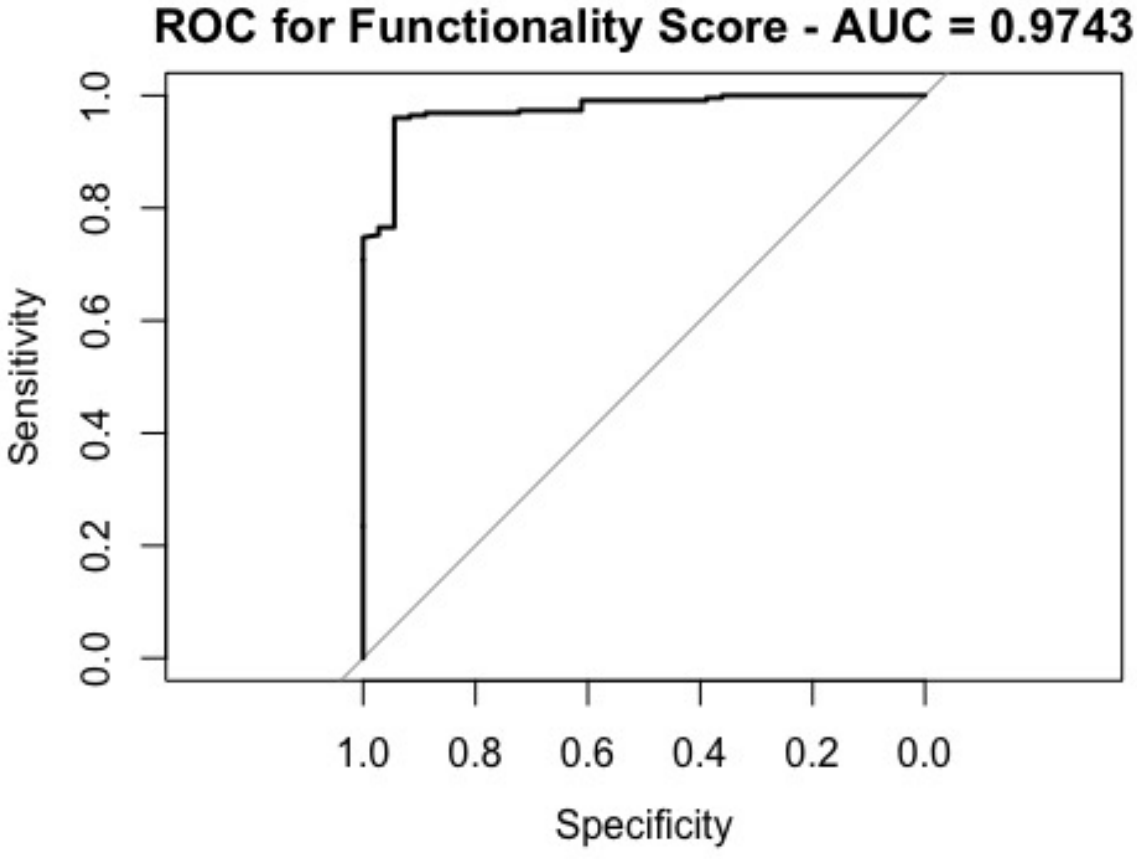
Posterior CDFs for Response



There is a stochastic ordering between outcome categories!
(p-value ≈ 0.035)

A Functionality Score

We compute the area under the curve as an individual functionality measure.



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.
- **52 Cell Subsets**
  - 25 CD4 cells.
  - 27 CD8 cells.

Analysis Goals

- **Problem:** We are interested in identifying response in Subsets X Protein pairs.
- **Solution:** Define each combination of Subset X Protein as a cell-subset.
  - Overall 184 subsets with non-negligible counts.
- Dependence structures should (and do!) sort themselves out.
- Include covariates?

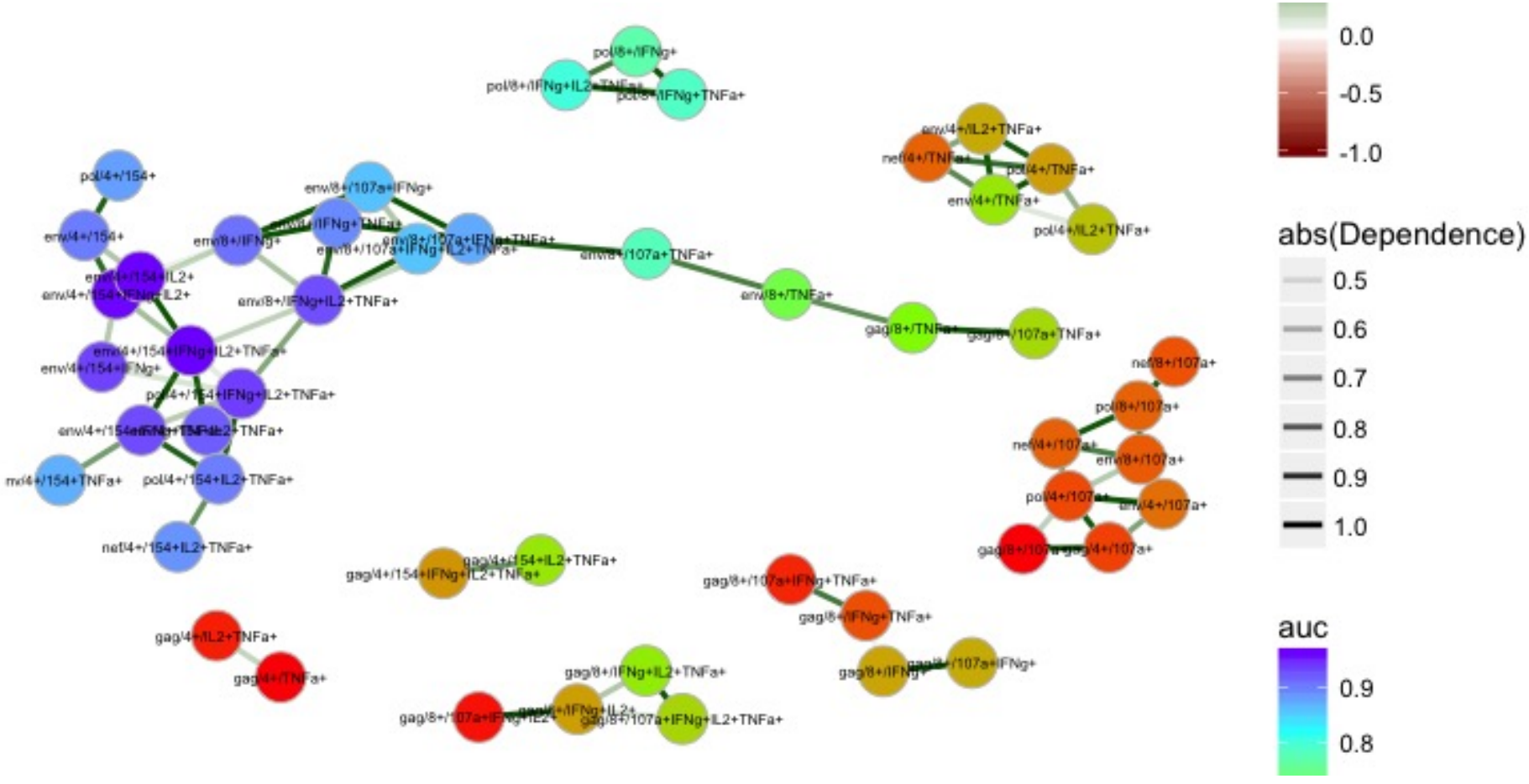
Analysis Goals

- **Problem:** We are interested in identifying response in Subsets X Protein pairs.
- **Solution:** Define each combination of Subset X Protein as a cell-subset.
  - Overall 184 subsets with non-negligible counts.
- Dependence structures should (and do!) sort themselves out.
- Include covariates?

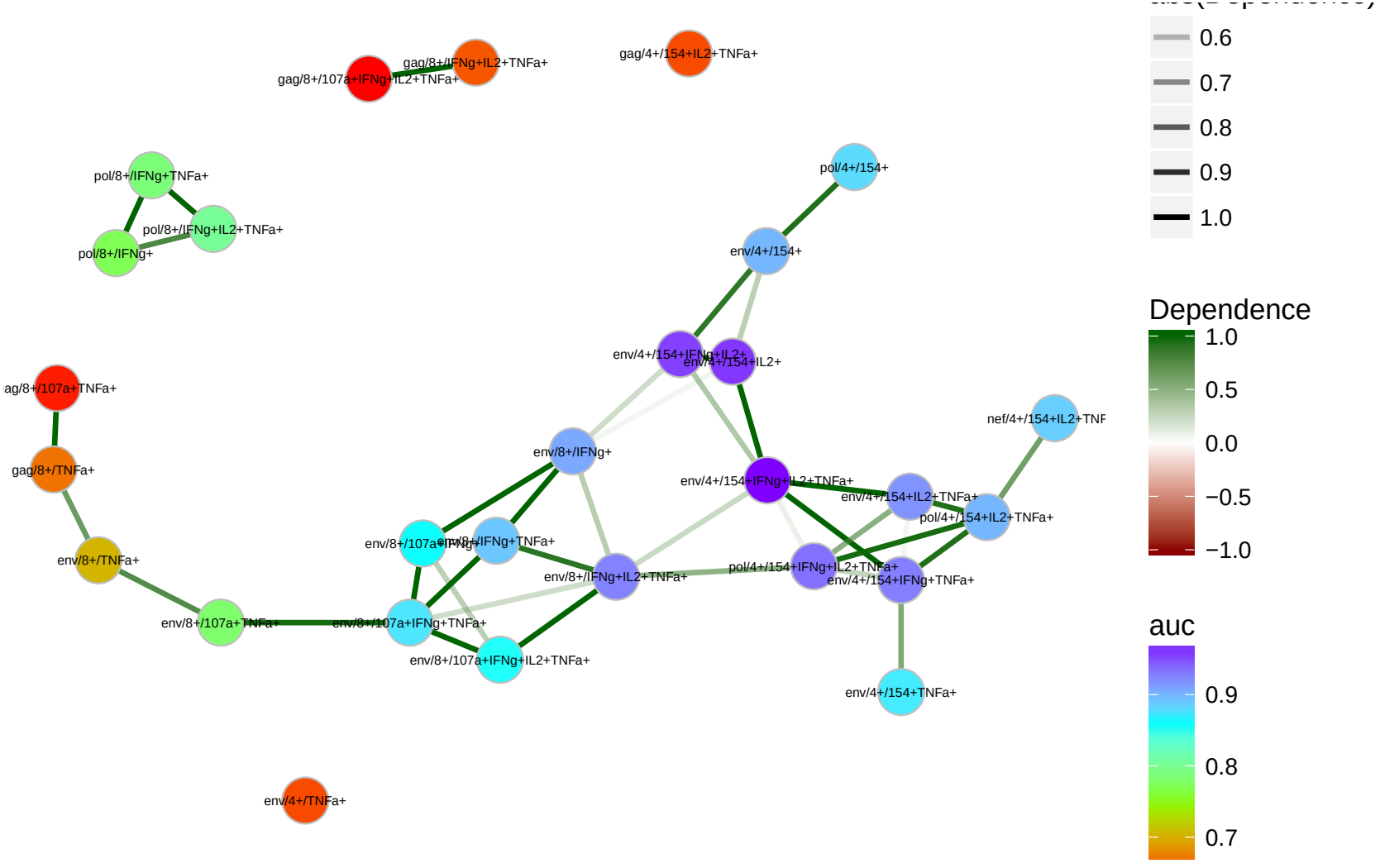
Analysis Goals

- **Problem:** We are interested in identifying response in Subsets X Protein pairs.
- **Solution:** Define each combination of Subset X Protein as a cell-subset.
  - Overall 184 subsets with non-negligible counts.
- Dependence structures should (and do!) sort themselves out.
- Include covariates?

Inferred Graph



Inferred Graph



Correlates for Infection-Status

As we have done for the RV144 dataset, we can correlate response in different cell-subsets with either **vaccination** status, or **infection** status.

Top Subsets For Vaccination Status (54 significant)

	subsets	aucs	pvals	qvals
14	env/4+/154+IFNg+IL2+TNFa+	0.9691178	1.464510e-35	2.694698e-33
16	env/4+/154+IL2+	0.9598316	4.338808e-33	7.983408e-31
13	env/4+/154+IFNg+IL2+	0.9561602	3.328590e-32	6.124605e-30
154	pol/4+/154+IFNg+IL2+TNFa+	0.9363460	4.651315e-28	8.558420e-26
12	env/4+/154+IFNg+	0.9337545	1.396123e-27	2.568867e-25
15	env/4+/154+IFNg+TNFa+	0.9287874	1.064343e-26	1.958392e-24

Top Subsets For Infection Status (4 significant)

	subsets	aucs	infectPvals	infectQvals
46	env/8+/IFNg+	0.7615854	6.594302e-06	0.001213352
31	env/8+/107a+IFNg+	0.7397561	3.645826e-05	0.006708319
50	env/8+/IL2+	0.7231707	1.192790e-04	0.021947336
47	env/8+/IFNg+IL2+	0.7204878	1.431531e-04	0.026340179
49	env/8+/IFNg+TNFa+	0.6943902	7.436270e-04	0.136827365

Aggregate Measures of Response

We have many more cell-subsets here, and can ask more interesting questions. Can we think of better aggregate measures?

- **Polyfunctionality:**
  - Polyfunctional cells produce multiple cytokines.
  - Few, but may play an important role in immunization.
  - **Implication:** Give higher weights to polyfunctional cells.

- **Breadth:**
  - How man stimulations does a subject respond to?
  - **Implication:** Give higher weights to first responsive subsets for a given stimulation.

Aggregate Measures of Response

We have many more cell-subsets here, and can ask more interesting questions. Can we think of better aggregate measures?

- **Polyfunctionality:**
  - Polyfunctional cells produce multiple cytokines.
  - Few, but may play an important role in immunization.
  - **Implication:** Give higher weights to polyfunctional cells.

- **Breadth:**
  - How man stimulations does a subject respond to?
  - **Implication:** Give higher weights to first responsive subsets for a given stimulation.

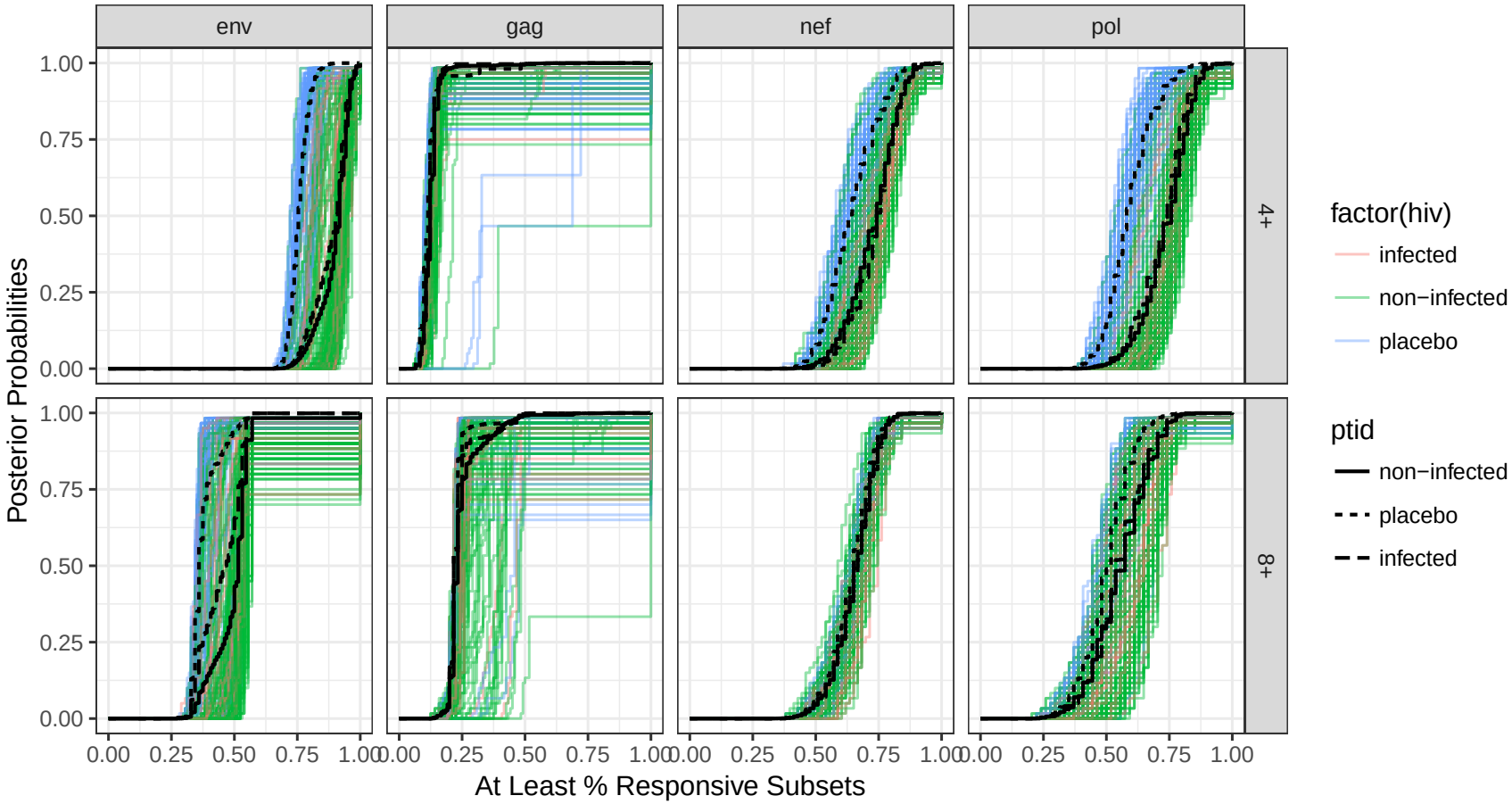
Aggregate Measures of Response

We have many more cell-subsets here, and can ask more interesting questions. Can we think of better aggregate measures?

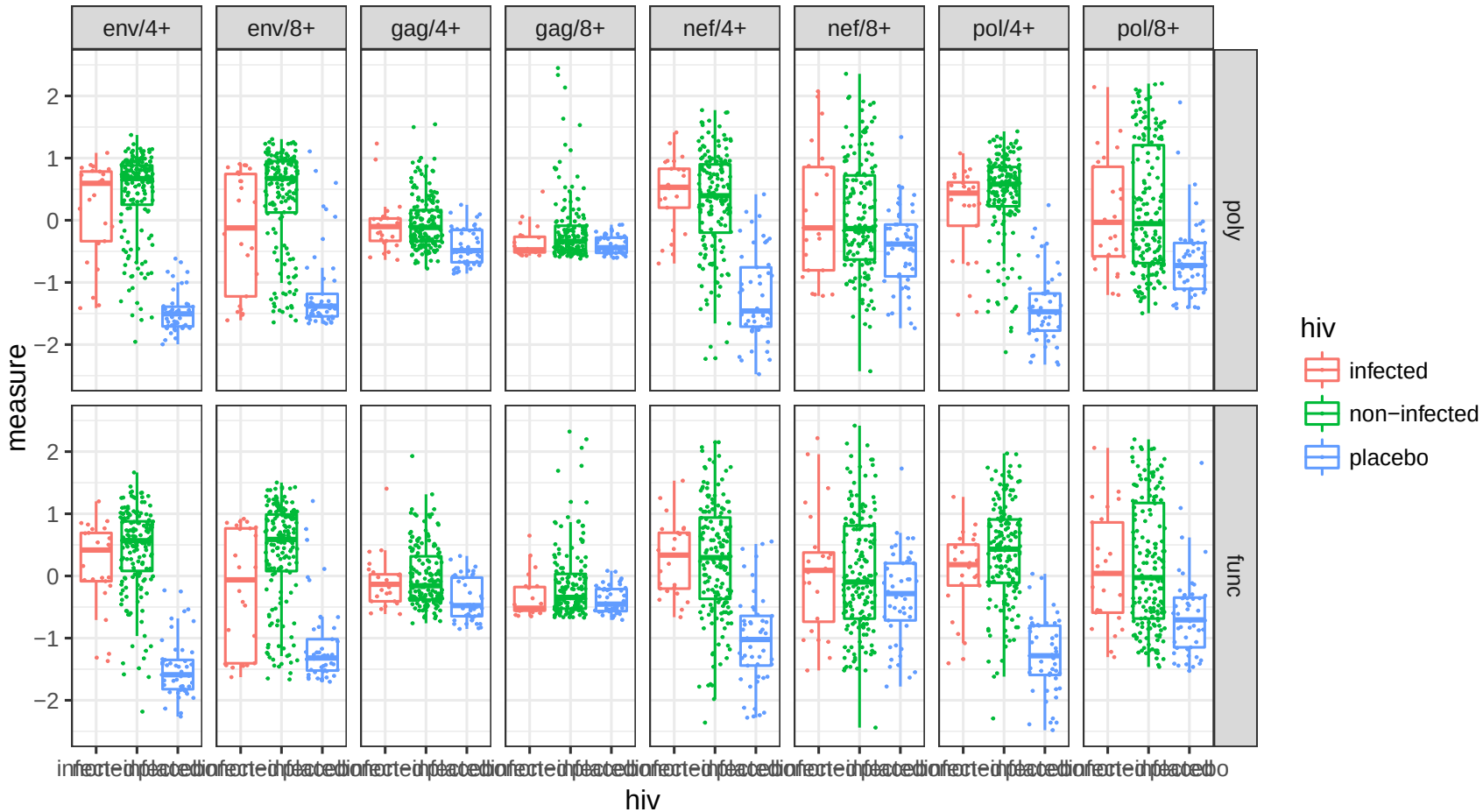
- **Polyfunctionality:**
  - Polyfunctional cells produce multiple cytokines.
  - Few, but may play an important role in immunization.
  - **Implication:** Give higher weights to polyfunctional cells.

- **Breadth:**
  - How man stimulations does a subject respond to?
  - **Implication:** Give higher weights to first responsive subsets for a given stimulation.

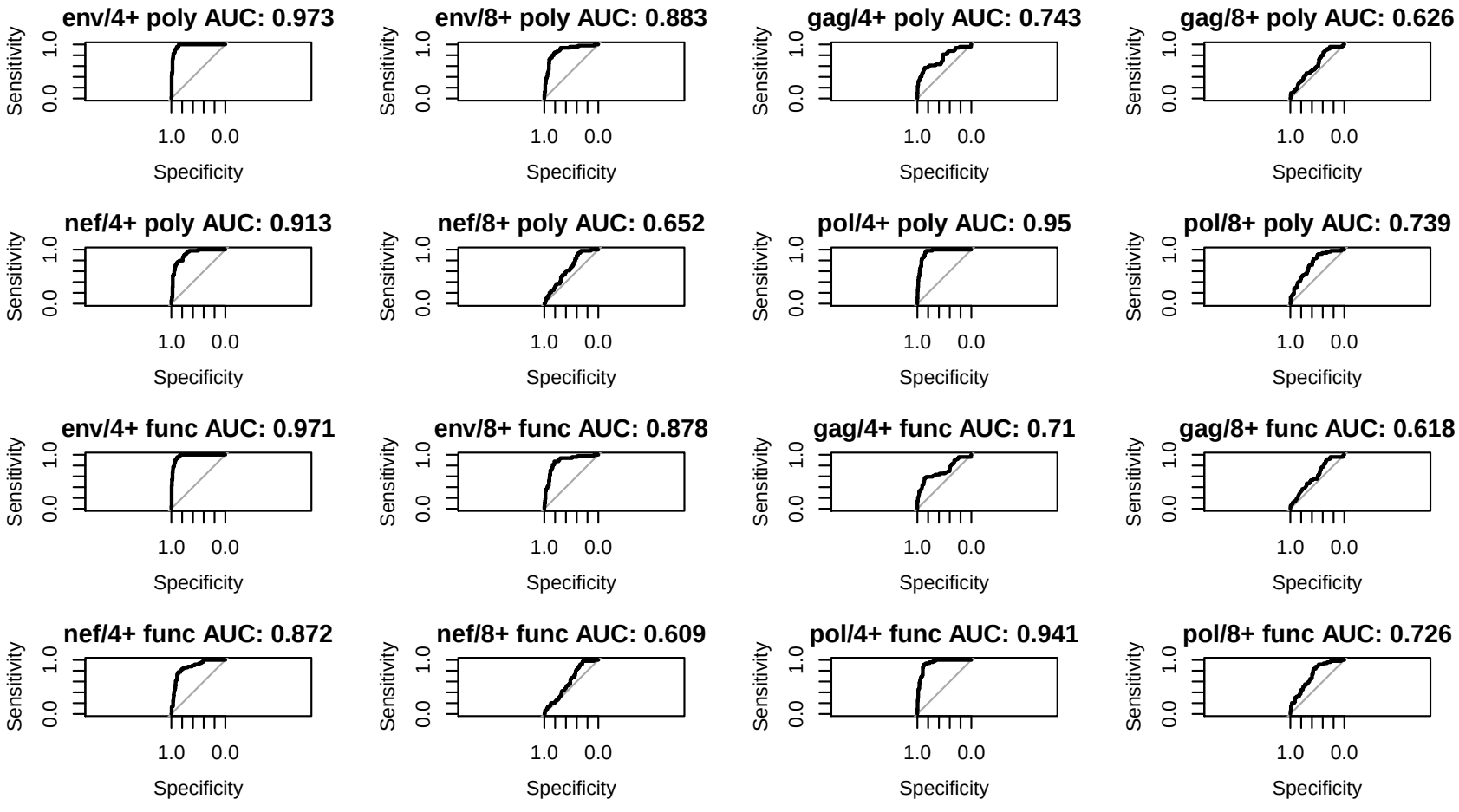
HVTN Polyfunctionality CDFs



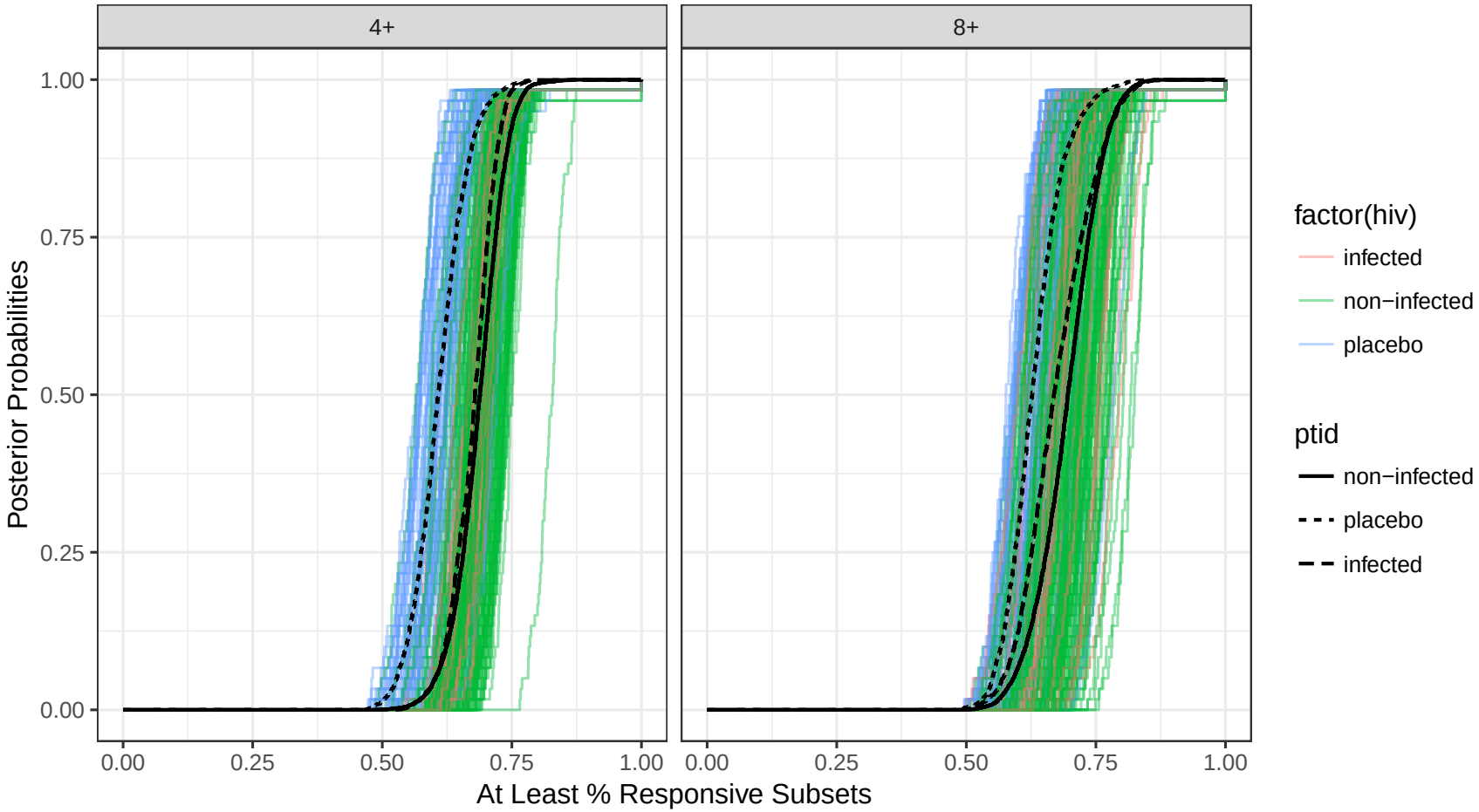
HVTN Polyfunctionality Score Boxplots



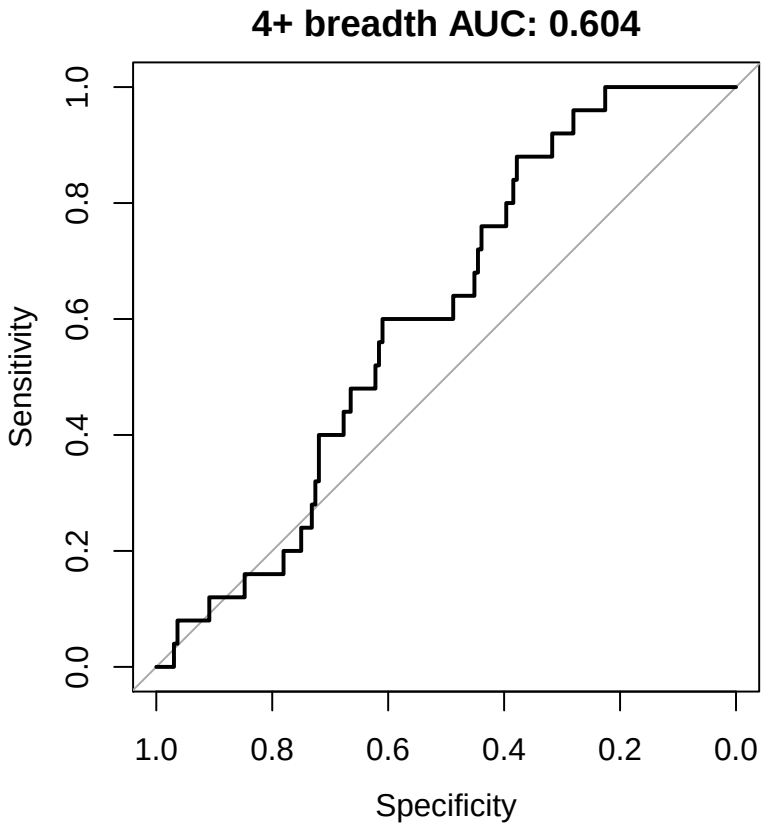
HVTN Polyfunctionality Score ROCs

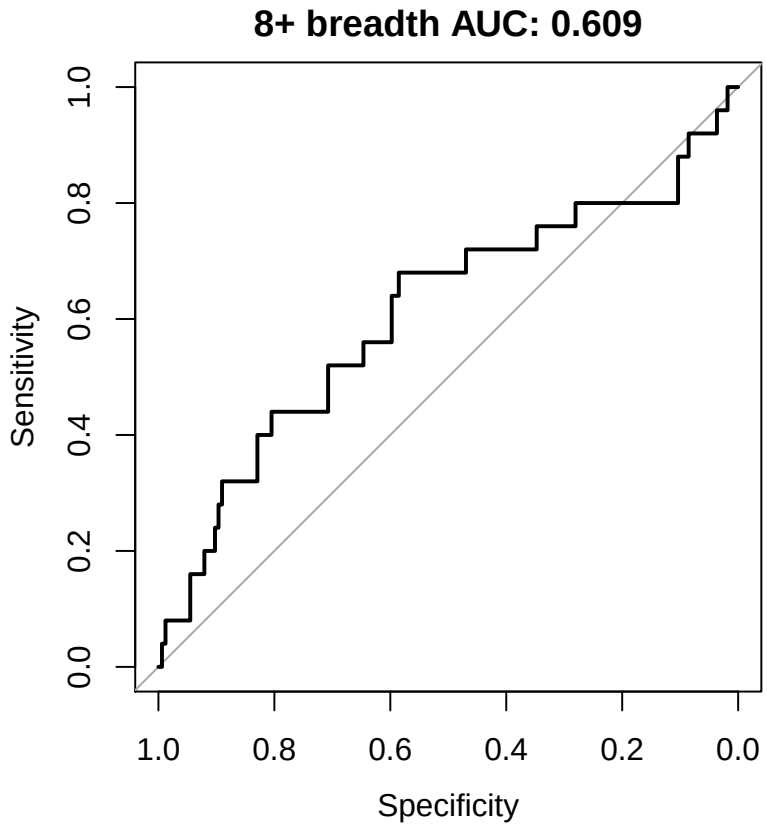


HVTN Breadth CDFs



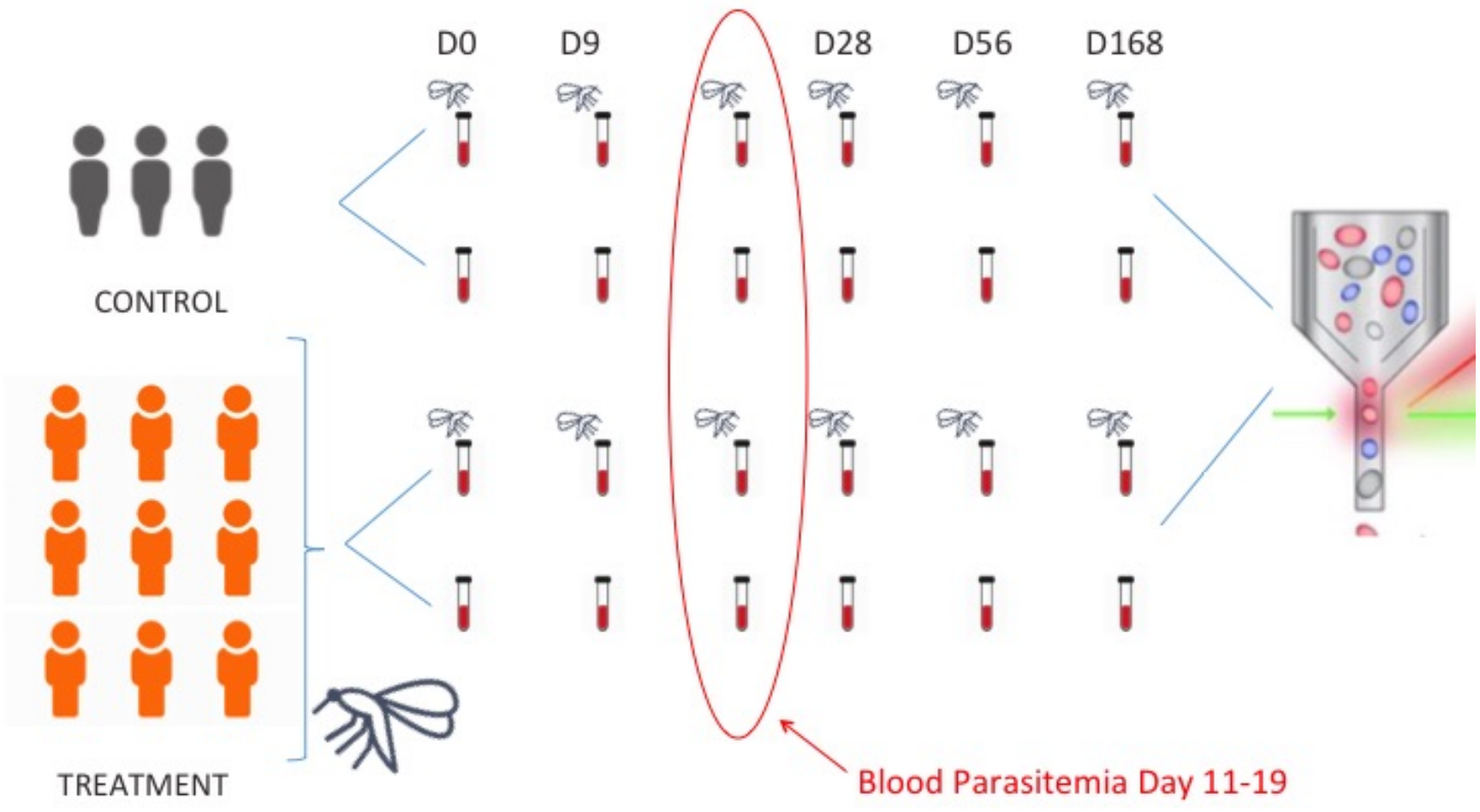
Breadth - ROC for Infection Study





P-value < 0.05 **for both ROCs**

Controlled Human Malaria Infection Study



Controlled Human Malaria Infection Study

- 9 subjects were infected with Malaria.
  - +3 controls.
- Blood samples were collected at 6 time points.
  - Day 0, day 9, blood parasitemia, Day 28, Day 56, Day 168.
- Two types of stimulation:
  - Infected/uninfected blood-cells.
- 53 cell subsets.
  - (10 types of cytokines in 8 cell-types)

Controlled Human Malaria Infection Study

- Individuals who experience malaria infections develop immunity.
  - All subject may exhibit response to stimulation.
  - Even at day 0!
  - What is the profile of the immune response?
- The immunity is not long lived.
  - We might expect to see a rise in response during experiment.
  - How fast does the response return to baseline?

Controlled Human Malaria Infection Study

- Individuals who experience malaria infections develop immunity.
  - All subject may exhibit response to stimulation.
  - Even at day 0!
  - What is the profile of the immune response?
- The immunity is not long lived.
  - We might expect to see a rise in response during experiment.
  - How fast does the response return to baseline?

Controlled Human Malaria Infection Study

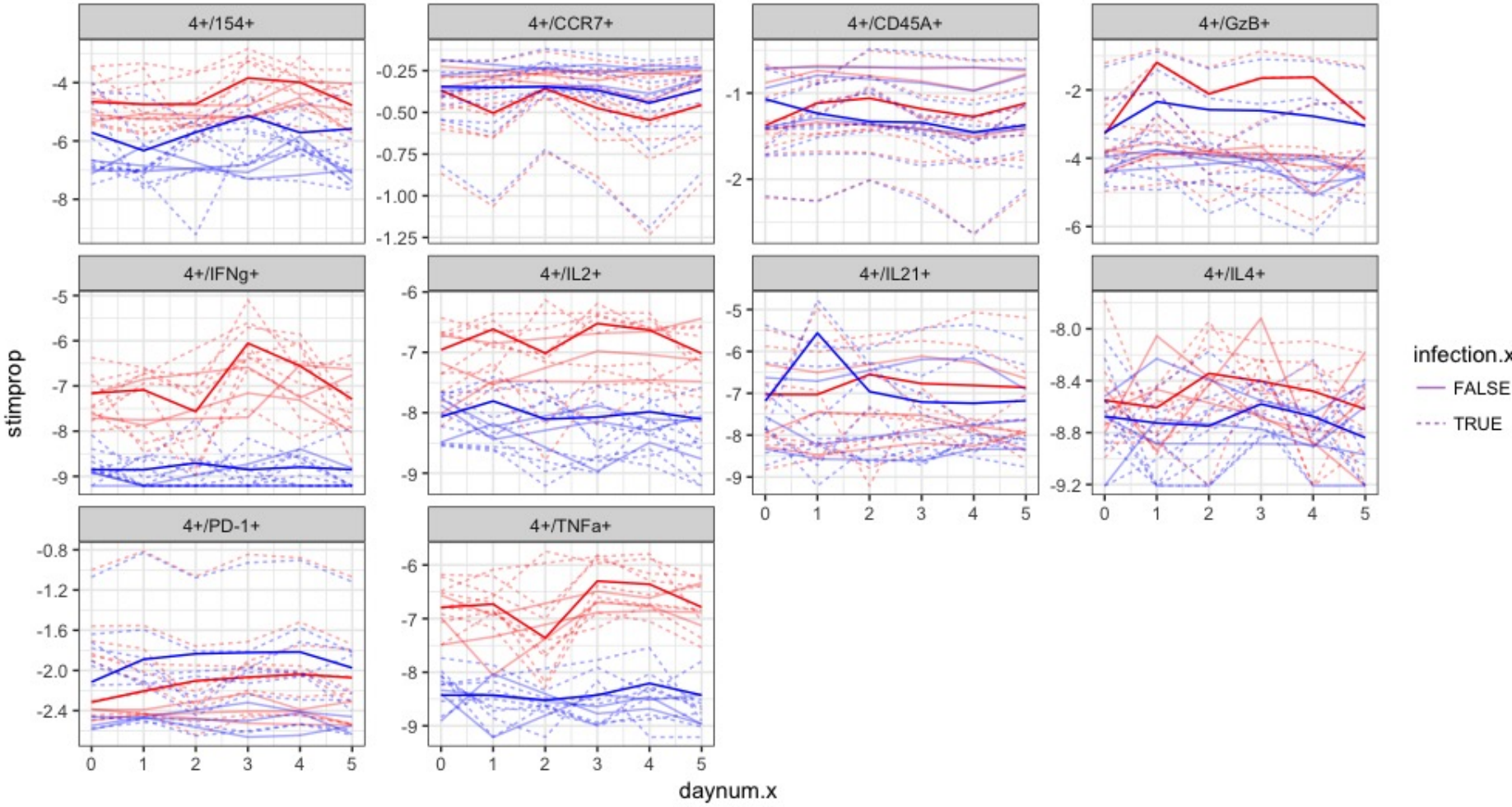


Figure: CD4 Helper Cells

FDR Adjusted p-values for CHMI Study

Standard errors for significance tests computed using Jackknife.

	4+	4+/CXCR5+	56+dim	56+hi	8+	8+/CXCR5+	NK T cells	PD-1+
154+	0.029	0.004			0.103	0.75	0.006	0.024
CCR7+	0.649	0.996			0.596	0.51		
CD45A+	0.575	0.307			0.543	0.54		
IFNg+	0.001	0.006	0.065	0.146	0.001		0.052	0.097
IL2+	0	0.005			0.119	0.56	0.321	0.052
IL21+	0.676	0.649	0.751	0.589	0.649		0.71	
IL4+	0.12	0.543		0.751	0.649		0.583	
TNFa+	0	0.001	0.261	0.309	0.276		0.053	0.09
GzB+	0.583		0.511	0.001	0.589		0.596	
PD-1+	0.751				0.596	0.83		

Controlled Human Malaria Infection Study

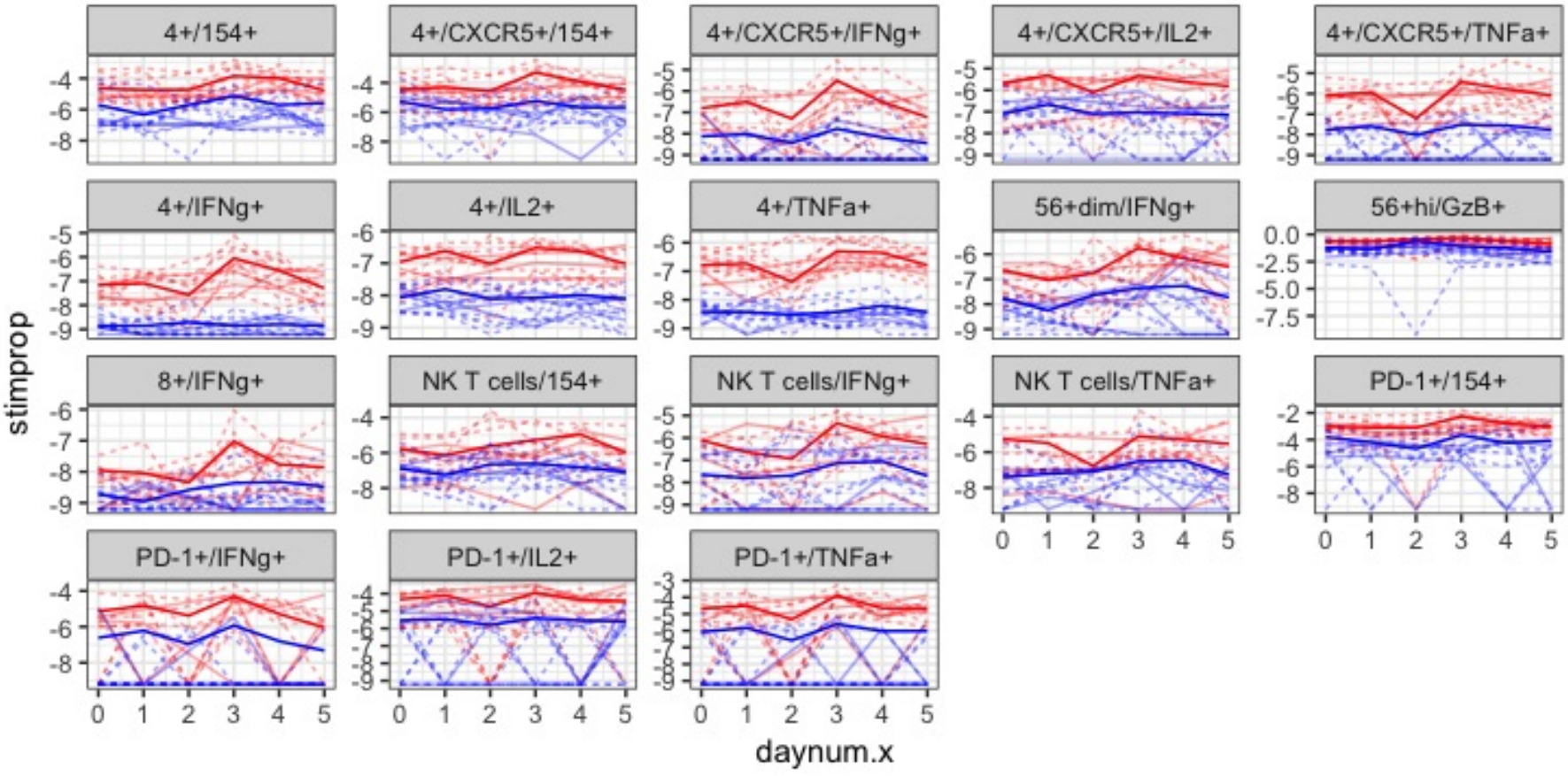


Figure: Significant Subsets

Enrichment Analysis and Valid Ad-Hoc Testing

The data seems to suggest testing for elevated response after parasitemia.

This effect may be small, and was identified based on the data.

Possible Solution:

- Test enrichments (groups of cells) to obtain more power.
- Perform a post-selection test for subsets within enrichments that pass a threshold.

We test:

- **Enrichments:** Th1, Th2, Gzb.
- **Hypotheses:** Overall effect, elevated response at 28, 56, 168.

Enrichment Analysis and Valid Ad-Hoc Testing

The data seems to suggest testing for elevated response after parasitemia.

This effect may be small, and was identified based on the data.

Possible Solution:

- Test enrichments (groups of cells) to obtain more power.
- Perform a post-selection test for subsets within enrichments that pass a threshold.

We test:

- **Enrichments:** Th1, Th2, Gzb.
- **Hypotheses:** Overall effect, elevated response at 28, 56, 168.

Enrichment Analysis and Valid Ad-Hoc Testing

The data seems to suggest testing for elevated response after parasitemia.

This effect may be small, and was identified based on the data.

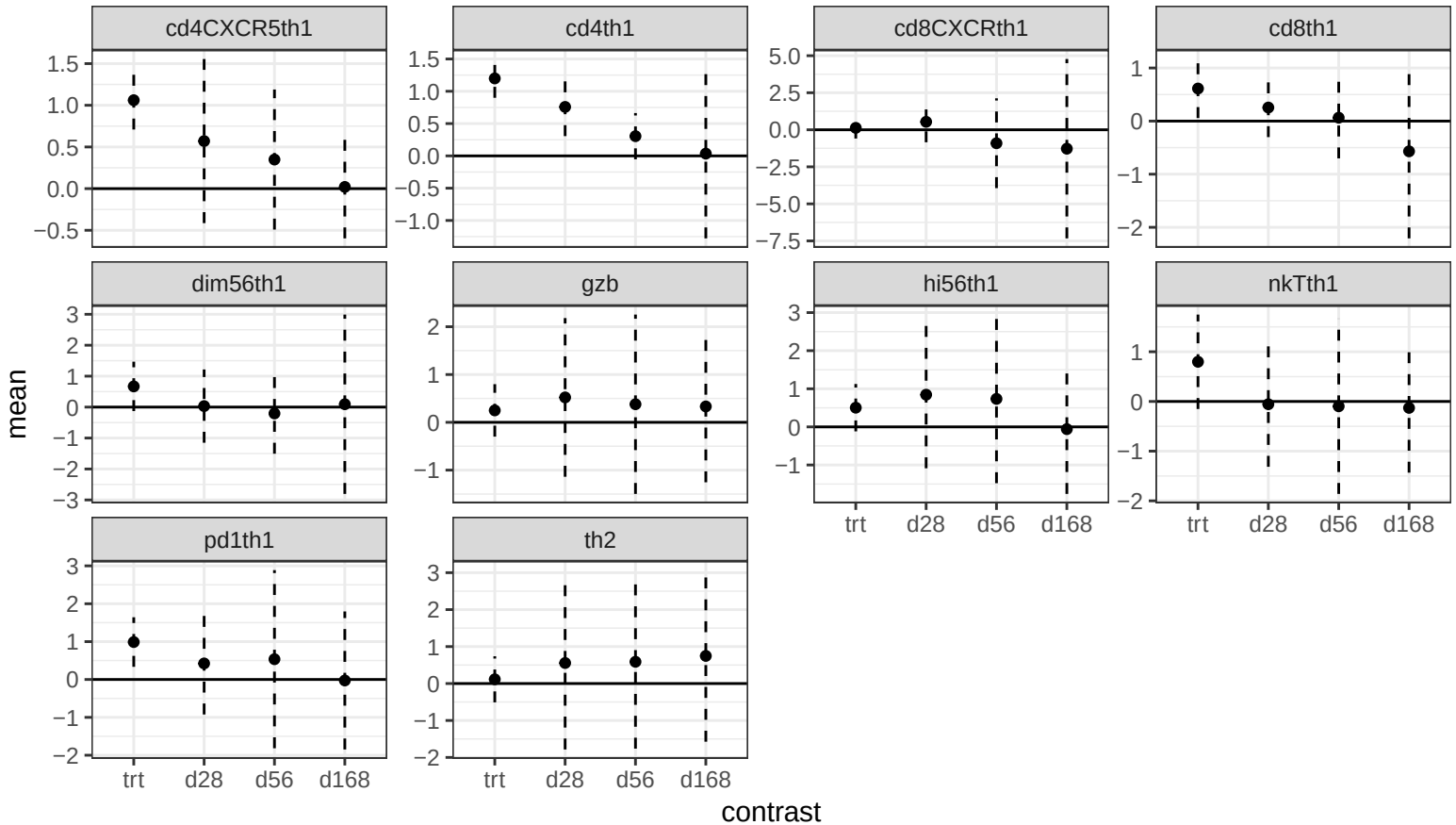
Possible Solution:

- Test enrichments (groups of cells) to obtain more power.
- Perform a post-selection test for subsets within enrichments that pass a threshold.

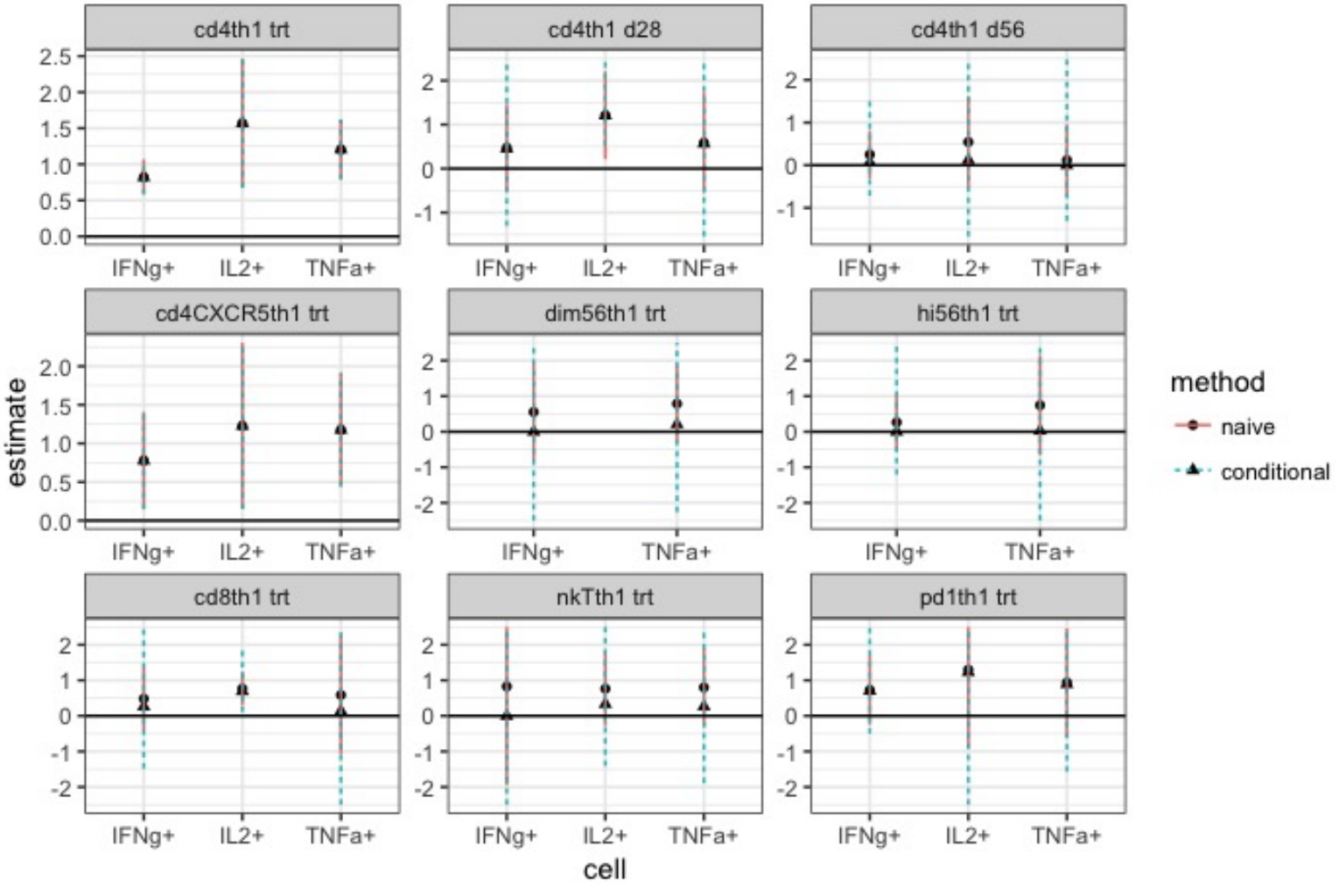
We test:

- **Enrichments:** Th1, Th2, Gzb.
- **Hypotheses:** Overall effect, elevated response at 28, 56, 168.

Controlled Human Malaria Infection Study



Controlled Human Malaria Infection Study



Thank you!

Questions?

AmitMeir@uw.edu