Analyzing Flow-Cytometry Count Data with Regression Mixtures

Amit Meir
University of Washington

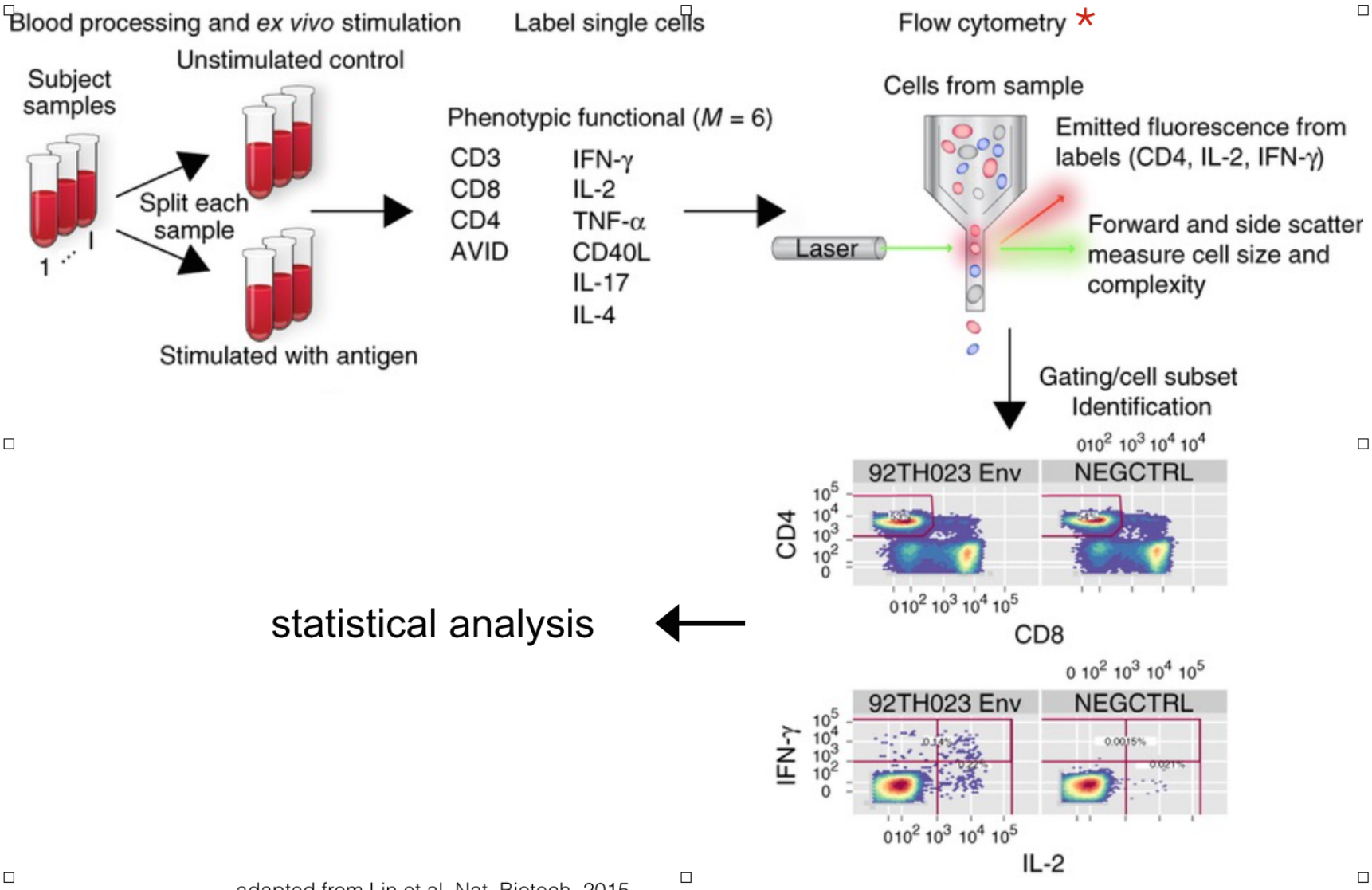
Joint work with
Raphael Gottardo and **Greg Finak**
Hutchinson Cancer Research Center

February 23, 2017

Outline

1. **Introduction** to Flow-Cytometry
2. **Motivation**
   - Existing methods
   - Why a regression model?
3. **Models**:
   - A Marginal Model
   - A Joint HMRF model
4. **Data analysis**
   - RV144 HIV Vaccine Trial
   - Controlled Human Malaria Infection Study.
5. **Computation** (time permitting)

The RV144 HIV Vaccine Trial



adapted from Lin et al, Nat. Biotech. 2015

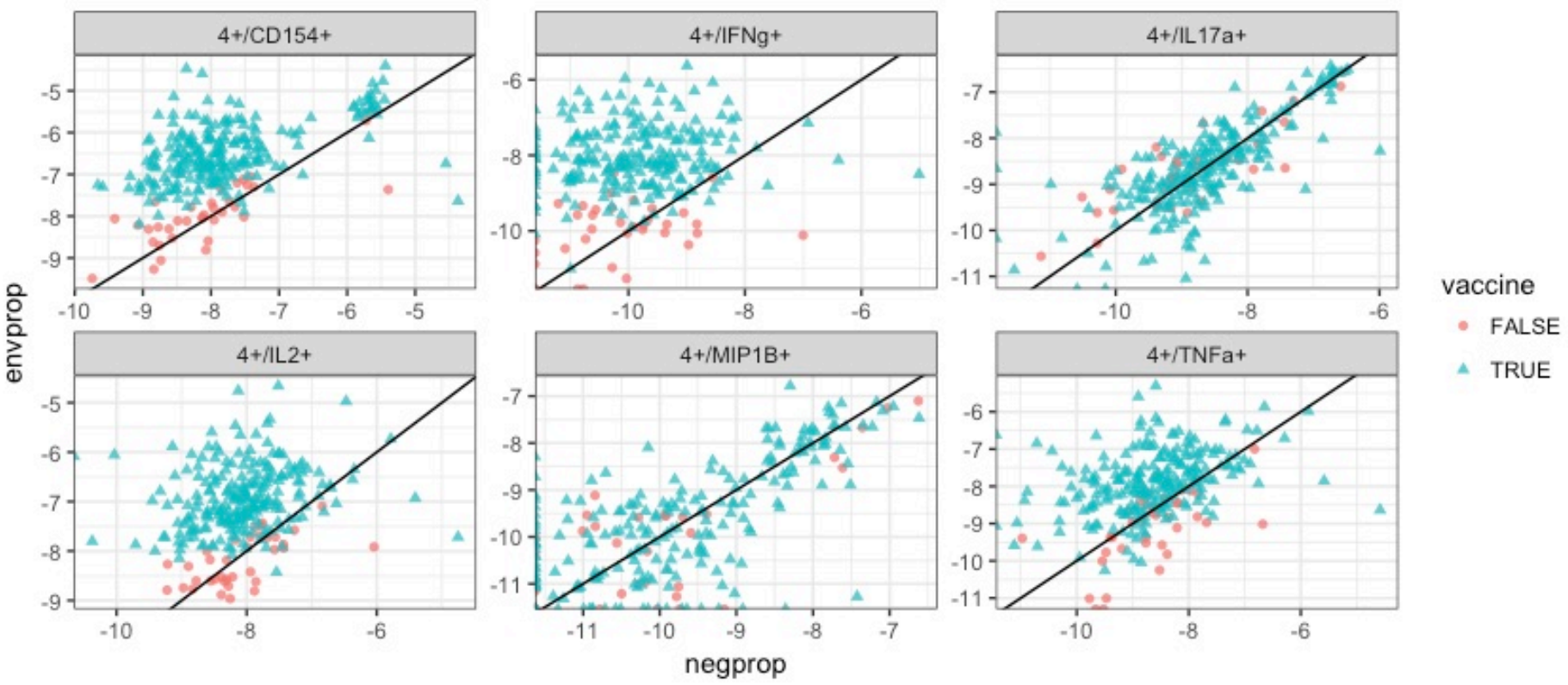
The RV144 HIV Vaccine Trial

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

The RV144 HIV Vaccine Trial

- **262 Subjects**
  - 226 Cases
  - 36 Controls
- **2 Types of stimulus**
  - HIV antigen
  - Negative control
- **6 types of cytokines.**

Marginal Counts for RV144



Motivation: COMPASS

ANALYSIS

_computational BIOLOGY

COMPASS identifies T-cell subsets correlated with clinical outcomes

Lin Lin[1], Greg Finak[1], Kevin Ushey[1], Chetan Seshadri[2], Thomas R Hawn[2], Nicole Frahm[1], Thomas J Scriba[3], Hassan Mahomed[3], Willem Hanekom[3], Pierre-Alexandre Bart[4], Giuseppe Pantaleo[4], Georgia D Tomaras[5], Supachai Rerks-Ngarm[6], Jaranit Kaewkungwal[7], Sorachai Nitayaphan[8], Punnee Pitisuttithum[9], Nelson L Michael[10], Jerome H Kim[10], Merlin L Robb[11], Robert J O'Connell[12], Nicos Karasavvas[12], Peter Gilbert[1], Stephen C De Rosa[1,13], M Juliana McElrath[1,2,13] & Raphael Gottardo[1]

Or in general, with immune response.

Motivation: COMPASS

ANALYSIS

_computational BIOLOGY

COMPASS identifies T-cell subsets correlated with clinical outcomes

Lin Lin[1], Greg Finak[1], Kevin Ushey[1], Chetan Seshadri[2], Thomas R Hawn[2], Nicole Frahm[1], Thomas J Scriba[3], Hassan Mahomed[3], Willem Hanekom[3], Pierre-Alexandre Bart[4], Giuseppe Pantaleo[4], Georgia D Tomaras[5], Supachai Rerks-Ngarm[6], Jaranit Kaewkungwal[7], Sorachai Nitayaphan[8], Punnee Pitisuttithum[9], Nelson L Michael[10], Jerome H Kim[10], Merlin L Robb[11], Robert J O'Connell[12], Nicos Karasavvas[12], Peter Gilbert[1], Stephen C De Rosa[1,13], M Juliana McElrath[1,2,13] & Raphael Gottardo[1]

Or in general, with immune response.

How do current methods work? (Approximately)

Current models are baseline/stimulation models.

- Unstimulated blood sample are compared stimulated ones.

For the unstimulated sample of the ith subject out of n, we sample a count proportion:

$$p_{i0} \sim \text{Dirichlet}(\alpha_0, \beta_0),$$

$$y_{i0} \sim \text{Multinomial}(N_{i0}, p_{i0}).$$

Let $k_i \in \{0,1\}^p$ indicate in which subsets i responds:

$$k_{ij} \sim \text{Ber}(w_j),$$

$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0}), \quad p_{i1,\tau=1}|p_{i0,\tau=0} \propto \text{Dirichlet}(\alpha_1, \beta_1)$$

$$y_{i1} \sim \text{Multinomial}(N_{i1}, p_{i1})$$

How do current methods work? (Approximately)

Current models are baseline/stimulation models.

- Unstimulated blood sample are compared stimulated ones.

For the unstimulated sample of the ith subject out of n, we sample a count proportion:

$$p_{i0} \sim \text{Dirichlet}(\alpha_0, \beta_0),$$

$$y_{i0} \sim \text{Multinomial}(N_{i0}, p_{i0}).$$

Let $k_i \in \{0,1\}^p$ indicate in which subsets i responds:

$$k_{ij} \sim \text{Ber}(w_j),$$

$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0}), \quad p_{i1,\tau=1}|p_{i0,\tau=0} \propto \text{Dirichlet}(\alpha_1, \beta_1)$$

$$y_{i1} \sim \text{Multinomial}(N_{i1}, p_{i1})$$

How do current methods work? (Approximately)

Current models are baseline/stimulation models.

- Unstimulated blood sample are compared stimulated ones.

For the unstimulated sample of the ith subject out of n, we sample a count proportion:

$$p_{i0} \sim \text{Dirichlet}(\alpha_0, \beta_0),$$

$$y_{i0} \sim \text{Multinomial}(N_{i0}, p_{i0}).$$

Let $k_i \in \{0, 1\}^p$ indicate in which subsets i responds:

$$k_{ij} \sim \text{Ber}(w_j),$$

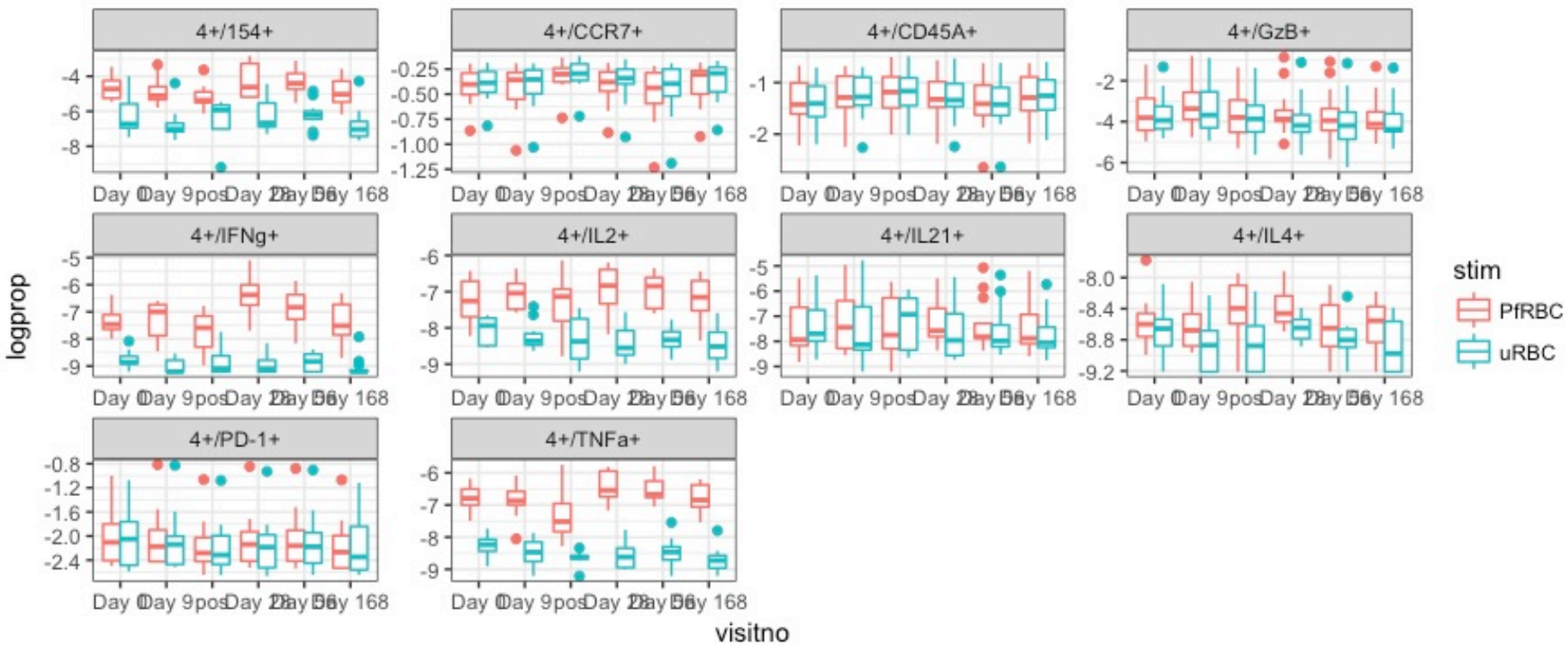
$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0}), \quad p_{i1,\tau=1} | p_{i0,\tau=0} \propto \text{Dirichlet}(\alpha_1, \beta_1)$$

$$y_{i1} \sim \text{Multinomial}(N_{i1}, p_{i1})$$

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



Motivation - A Regression Model

- We want to be able to include covariates:
  - Batch effects.
  - Other covariates such as age, gender...
- Longitudal data.
- More than one stimulation.
- Explicit dependence model:
  - For the observed proportions.
  - For response/non-response.

Motivation - A Regression Model

- We want to be able to include covariates:
  - Batch effects.
  - Other covariates such as age, gender...
- Longitudal data.
- More than one stimulation.
- Explicit dependence model:
  - For the observed proportions.
  - For response/non-response.

Motivation - A Regression Model

- We want to be able to include covariates:
  - Batch effects.
  - Other covariates such as age, gender...
- Longitudal data.
- More than one stimulation.
- Explicit dependence model:
  - For the observed proportions.
  - For response/non-response.

Motivation - Unique Challanges

- **Dependence**
  - Within sample between cell subsets.
  - Within subject / across time.
- **Heterogenous treatment effect**
- **Over-dispersed Binomial counts**

A Marginal Model - Single Subset

Indexing: **i**-subject, **t**- stimulation/time-point.

Overdispersion $\Rightarrow$ Beta Binomial Model.

$$\text{logit}(\mu_{it}) = X_{it}\beta,$$

$$p_{it} \sim \text{Beta}(M\mu, M(1-\mu)),$$

$$y_{it} \sim \text{Binom}(N_{it}, p_{it})$$

A Marginal Model - Single Subset

Indexing: **i**-subject, **t**- stimulation/time-point.

Dependence $\Rightarrow$ 'random' subject baseline:

$$\nu_i \sim N(0, \sigma^2)$$

$$\text{logit}(\mu_{it}) = X_{it}\beta + \nu_i$$

$$p_{it} \sim \text{Beta}(M\mu, M(1-\mu)),$$

$$y_{it} \sim \text{Binom}(N_{it}, p_{it})$$

A Marginal Model - Single Subset

Indexing: **i**-subject, **t**- stimulation, **k**- cluster.

Non-response $\Rightarrow$ Mixture-Model:

$$k \sim Ber(\theta),$$

$$\text{logit}(\mu_{itk}) = X_{it}\beta + T_{it}\tau_k + \nu_i,$$

- T a matrix of covariates related to the treatment.
- τ_k equals 0 if $k = 0$ or $\tau \neq 0$ if $k = 1$.

A Marginal Model - Recap

Indexing: **i**-subject, **t**- stimulation, **k**- cluster.

$$\nu_i \sim N(0, \sigma^2),$$

$$k \sim Ber(\theta),$$

$$\text{logit}(\mu_{itk}) = X_{it}\beta + T_{it}\tau_k + \nu_i,$$

$$p_{it} \sim \text{Beta}(M\mu, M(1-\mu)),$$

$$y_{it} \sim \text{Binom}(N_{it}, p_{it})$$

Model can be estimated via an EM algorithm

Wellness of Fit Evaluation

How do we evaluate the model?

- We fit the model without information regarding the true treatment allocation.
- The model should be able to discriminate between vaccinees and placebos.
- We use three type of figures:
  - Scatter plots w/classification information.
  - Receiver-Operator Curves.
  - False Detection Rates.

Wellness of Fit Evaluation

How do we evaluate the model?

- We fit the model without information regarding the true treatment allocation.
- The model should be able to discriminate between vaccinees and placebos.
- We use three type of figures:
  - Scatter plots w/classification information.
  - Receiver-Operator Curves.
  - False Detection Rates.

Marginal Model - Results

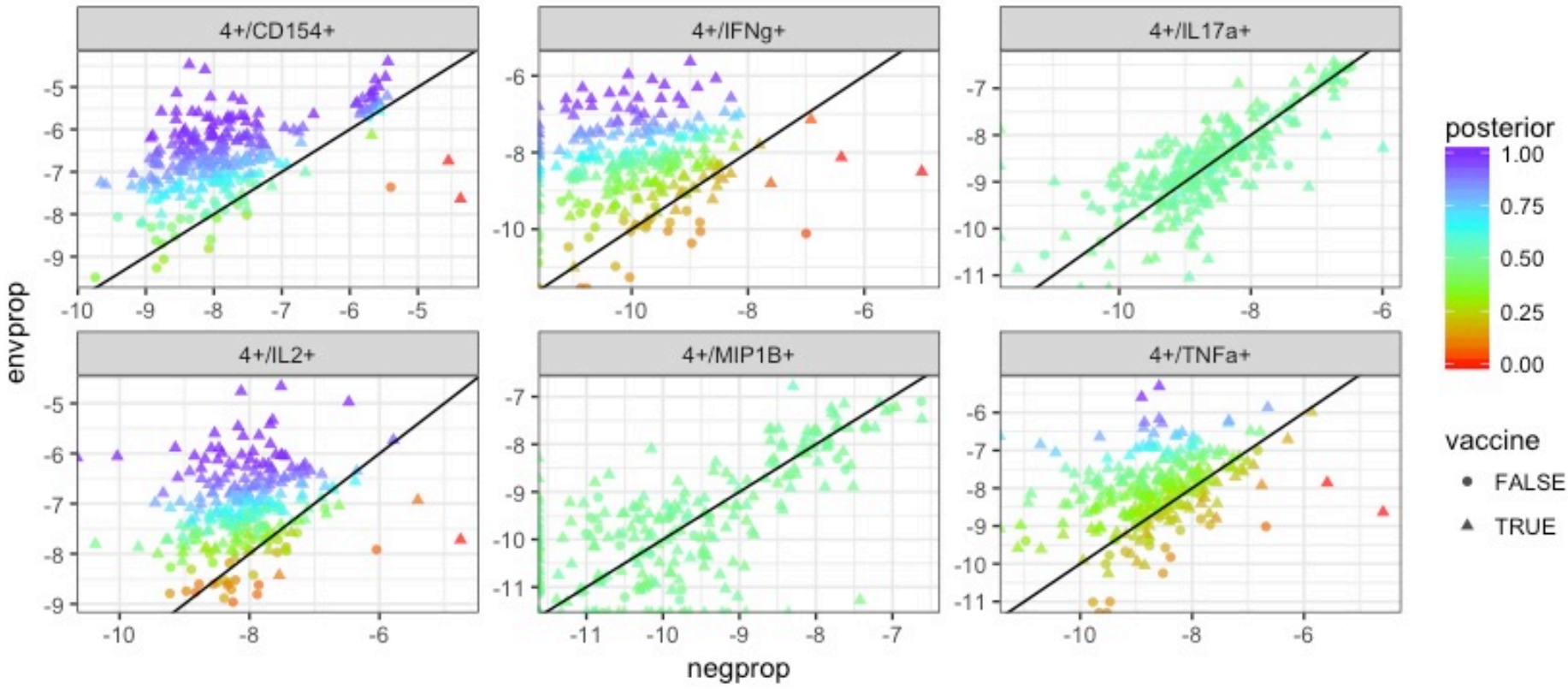


Figure: Posterior Probabilities for RV144 dataset - Independence Model

Comparison w/ MIMOSA - Finak et al. (2013)

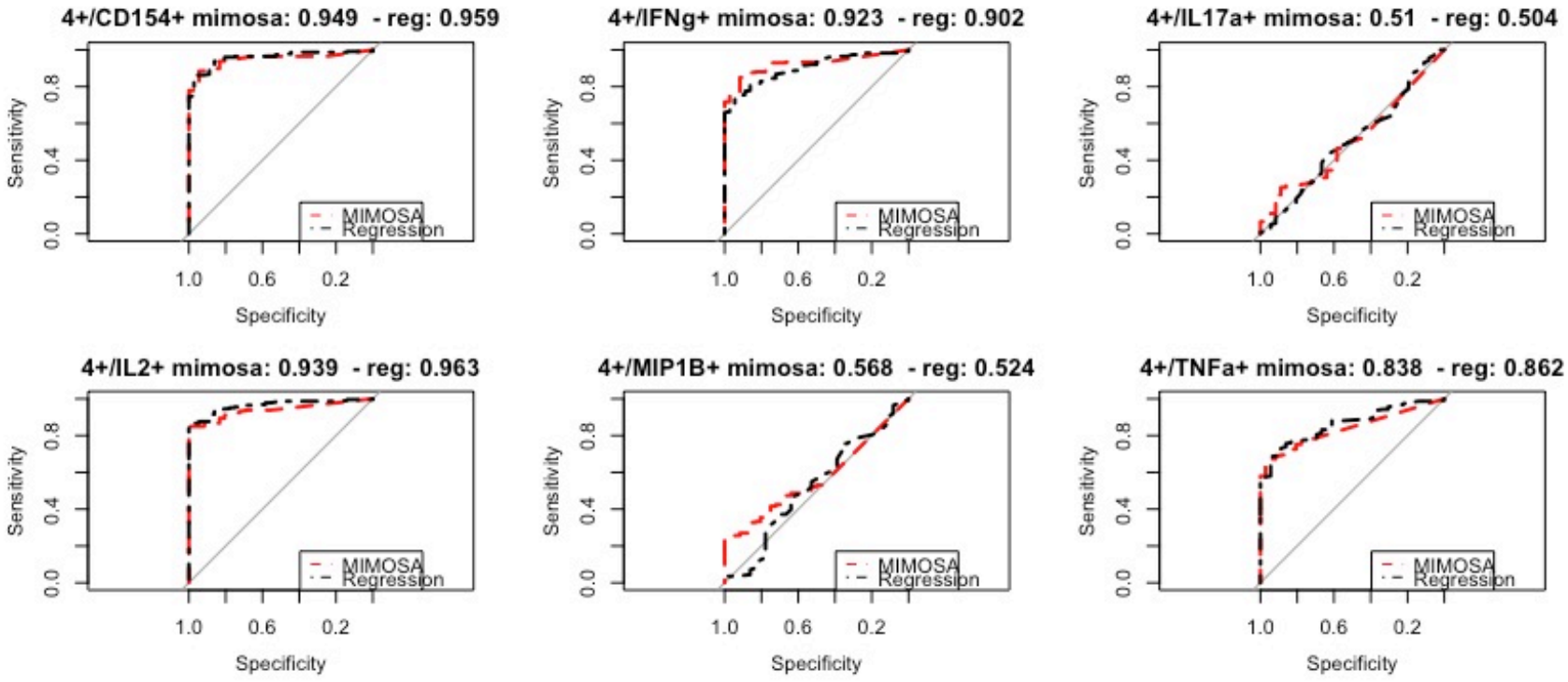


Figure: Comparison with MIMOSA (univariate COMPASS)

Comparison w/ MIMOSA - Finak et al. (2013)

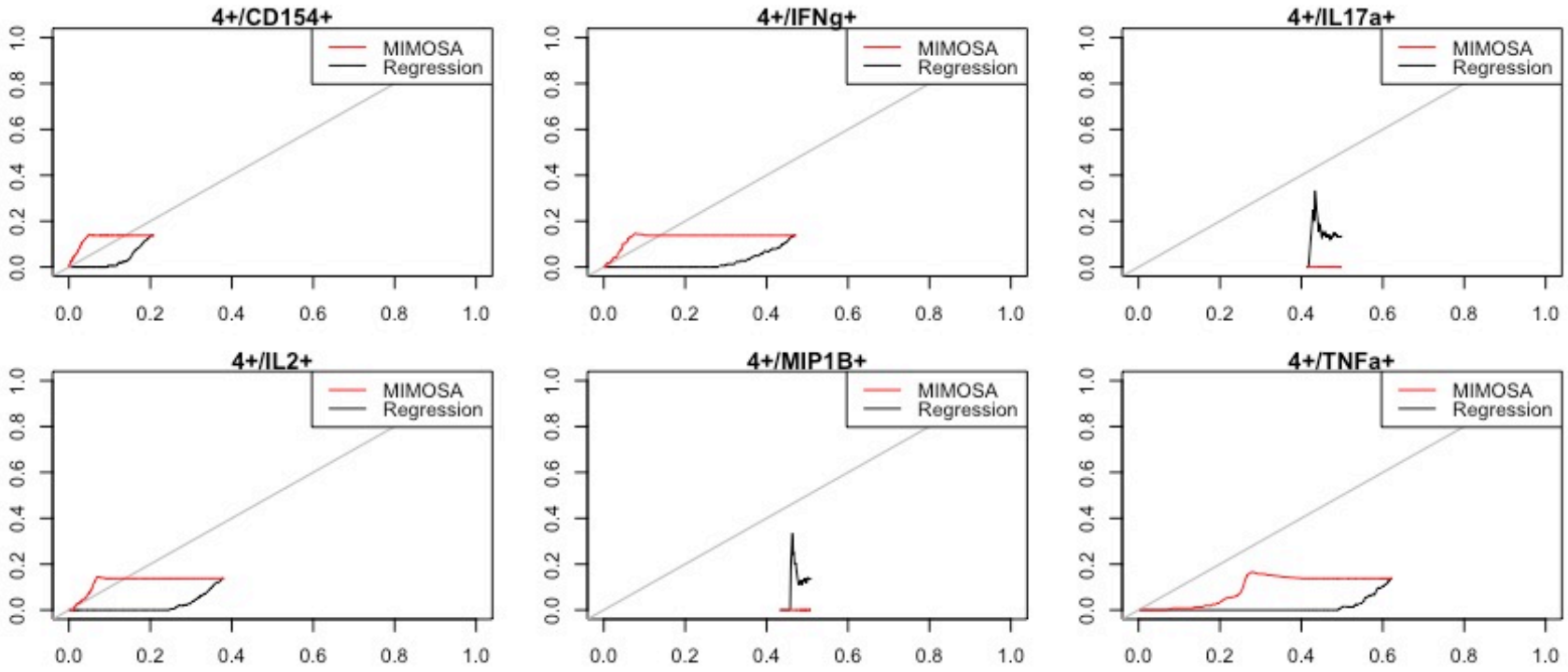


Figure: Comparison with MIMOSA (univariate COMPASS)

Subject-Response Model

Performing analysis for each cell-subset at a time doesn't use all of the information available.

- Random Effects are correlated, can be estimated better simultaneously.
- Correlation structure might be of interest in itself.
- Response is probably not independent across cell-subsets.
- We might be able to improve classification of response by looking at several cell-subsets at once.

Subject-Response Model

Performing analysis for each cell-subset at a time doesn't use all of the information available.

- Random Effects are correlated, can be estimated better simultaneously.
- Correlation structure might be of interest in itself.
- Response is probably not independent across cell-subsets.
- We might be able to improve classification of response by looking at several cell-subsets at once.

Subject-Response Model

Performing analysis for each cell-subset at a time doesn't use all of the information available.

- Random Effects are correlated, can be estimated better simultaneously.
- Correlation structure might be of interest in itself.
- Response is probably not independent across cell-subsets.
- We might be able to improve classification of response by looking at several cell-subsets at once.

A Hidden Markov Random Field Model

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

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

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

$$P(k) \propto \sum_{j=1}^{p} k_j \theta_j + \sum_{s \neq t} k_t k_s \theta_{st},$$

$$P(k_j = 1 | k_{-j}) = \theta_j + \sum_{t \neq j} k_t \theta_{tj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Hidden Markov Random Field Model

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

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

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

$$P(k) \propto \sum_{j=1}^{p} k_j \theta_j + \sum_{s \neq t} k_t k_s \theta_{st},$$

$$P(k_j = 1 | k_{-j}) = \theta_j + \sum_{t \neq j} k_t \theta_{tj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Hidden Markov Random Field Model

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

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

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

$$P(k) \propto \sum_{j=1}^{p} k_j \theta_j + \sum_{s \neq t} k_t k_s \theta_{st},$$

$$P(k_j = 1 | k_{-j}) = \theta_j + \sum_{t \neq j} k_t \theta_{tj}.$$

We can induce sparsity through an ℓ_1 penalty.

A Hidden Markov Random Field Model

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

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

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

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

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

Marginal Model - Results

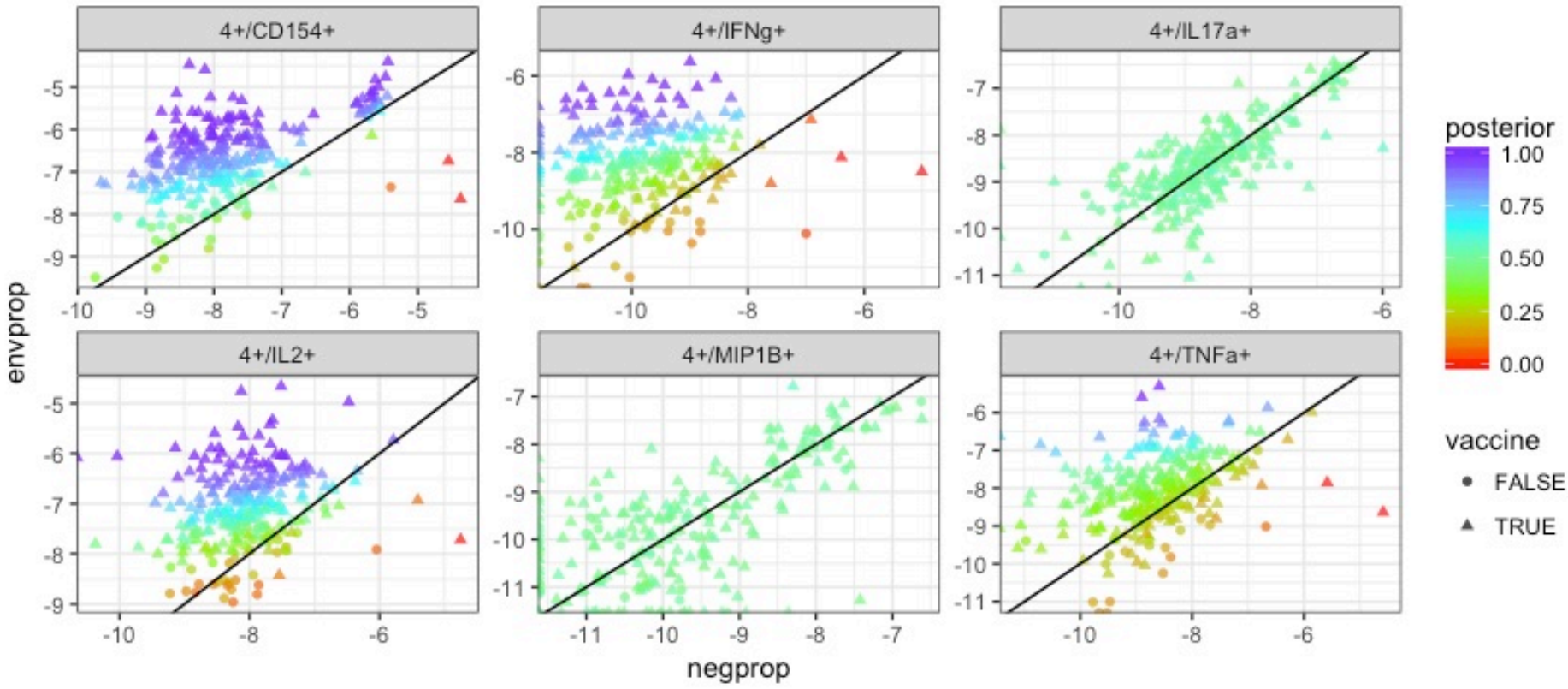


Figure: Posterior Probabilities for RV144 dataset - Independence Model

HMRF Model - Results

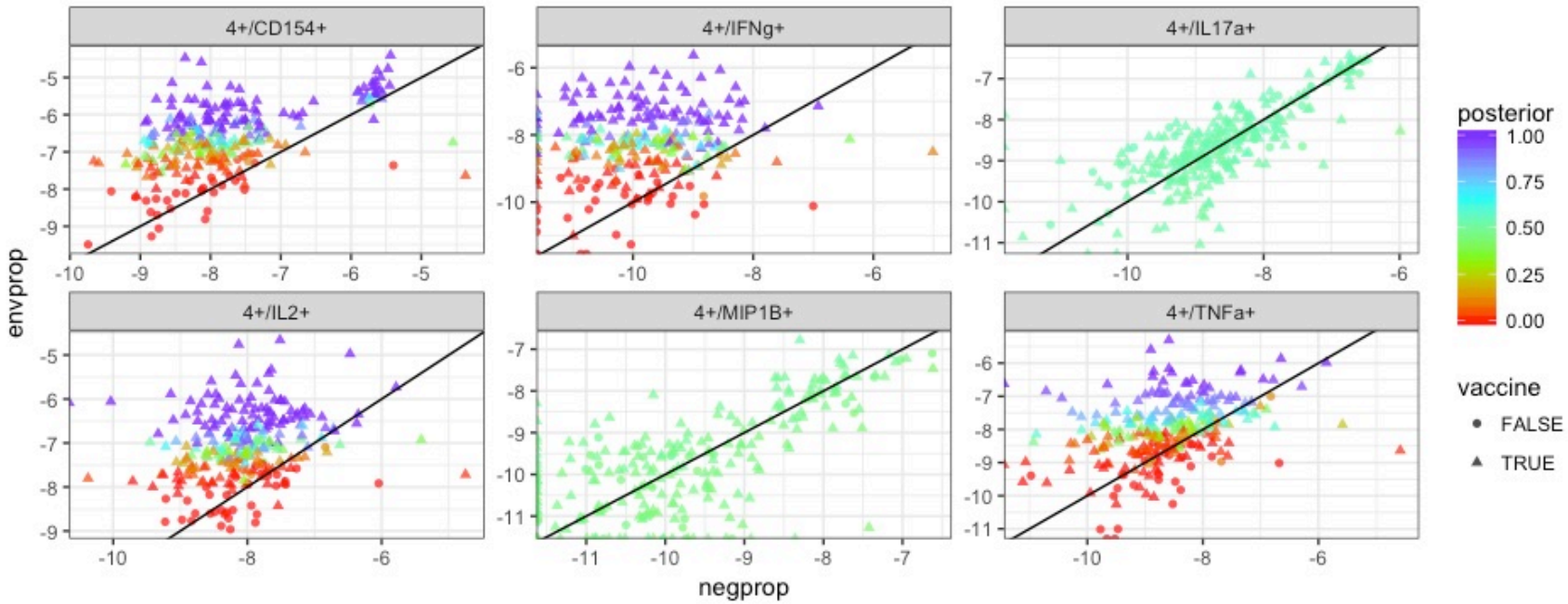


Figure: Scatter Plot for HMRF Modle Model

Subset-Response Model - Results

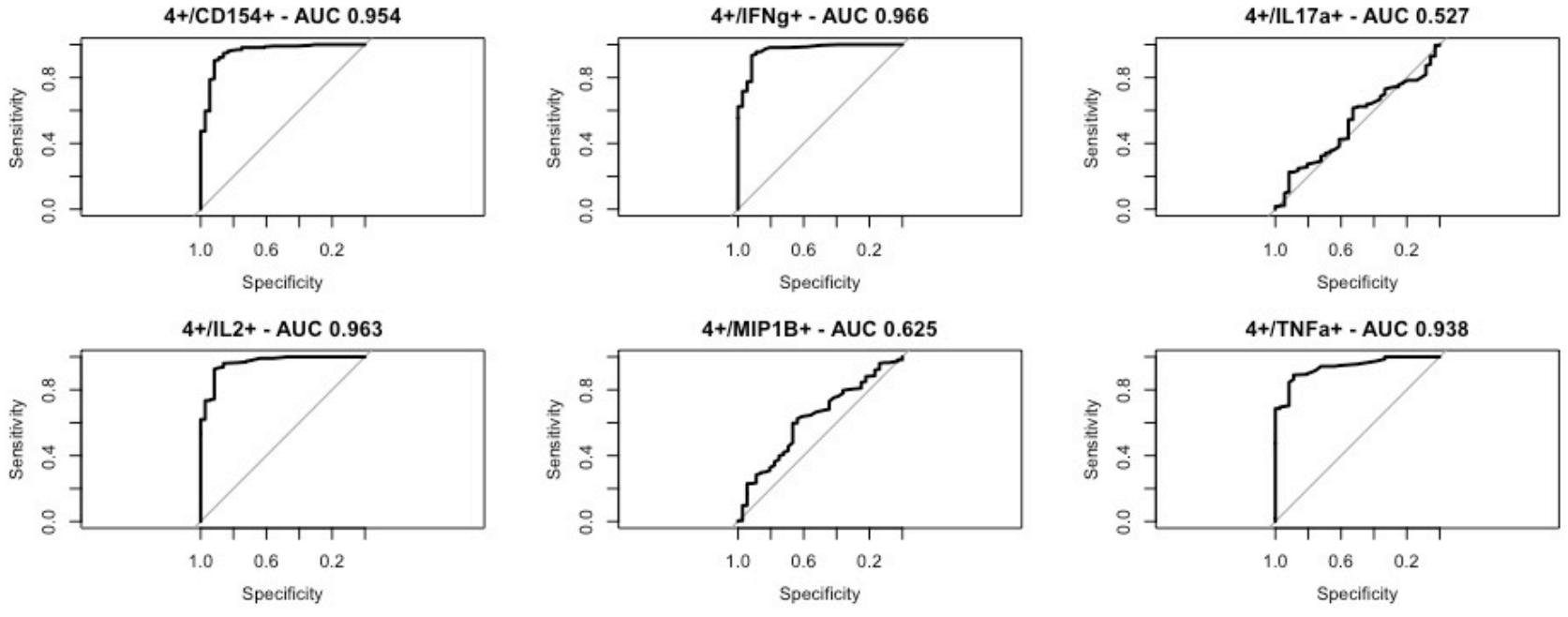


Figure: ROC for HMRF Modle

Subset-Response Model - Results

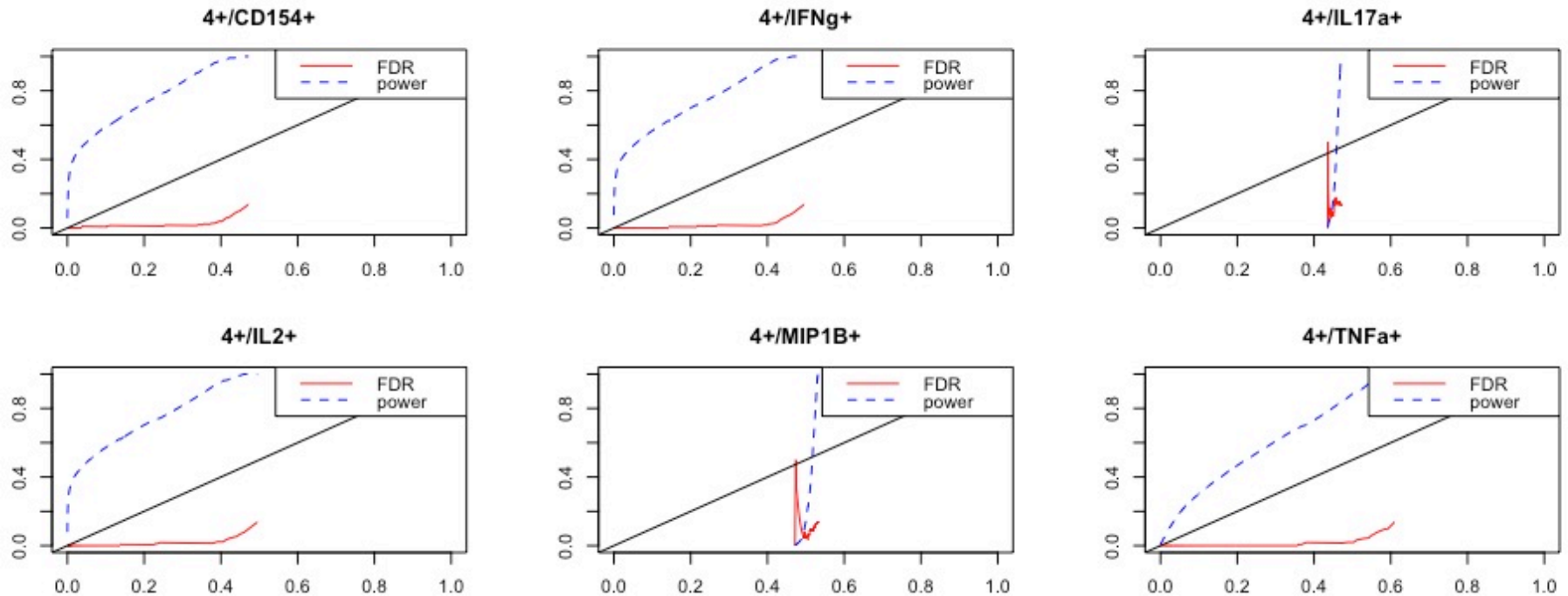


Figure: FDR for HMRF Modle

Simulated Data Experiment

- Posterior probabilities are not well calibrated.
  - Might be due to true non-response.
- Is the optimization algorithm estimating the model properly?
- Does the model fit the data well?
- To find out:
  - We generate data according to the estimated model.
  - Fit should be perfect.
  - Is the artificial data similar to the real data?

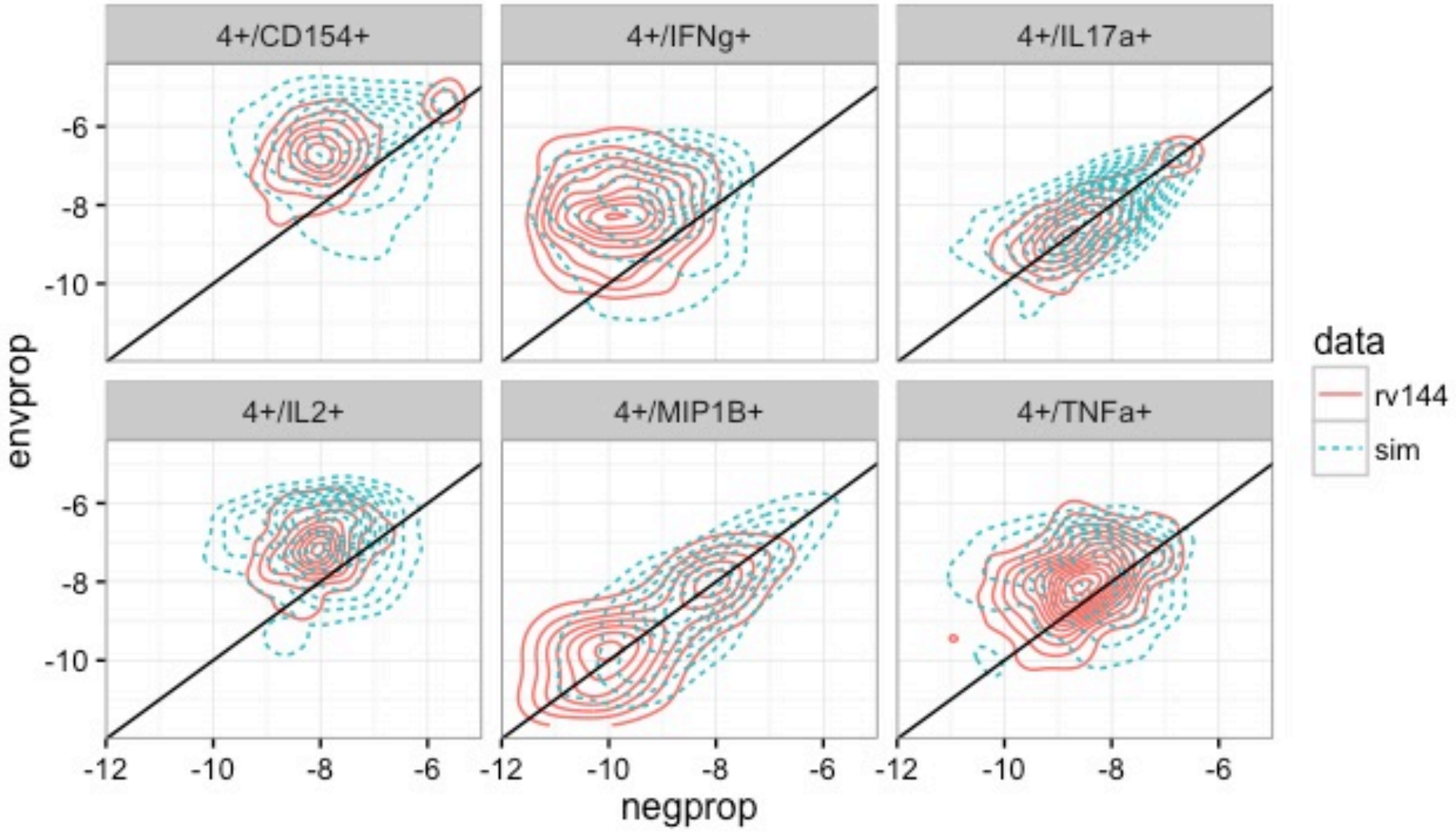
Simulated Data Experiment

- Posterior probabilities are not well calibrated.
  - Might be due to true non-response.
- Is the optimization algorithm estimating the model properly?
- Does the model fit the data well?
- To find out:
  - We generate data according to the estimated model.
  - Fit should be perfect.
  - Is the artificial data similar to the real data?

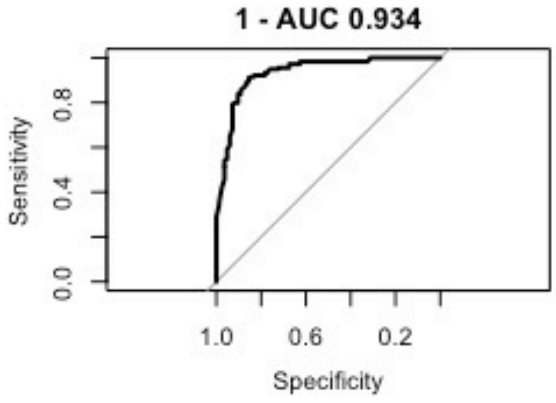
Simulated Data Experiment

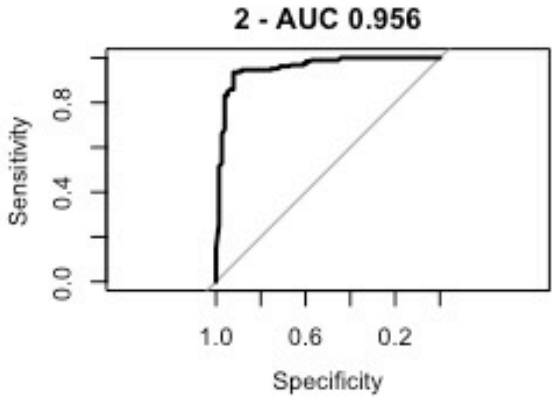
- Posterior probabilities are not well calibrated.
  - Might be due to true non-response.
- Is the optimization algorithm estimating the model properly?
- Does the model fit the data well?
- To find out:
  - We generate data according to the estimated model.
  - Fit should be perfect.
  - Is the artificial data similar to the real data?

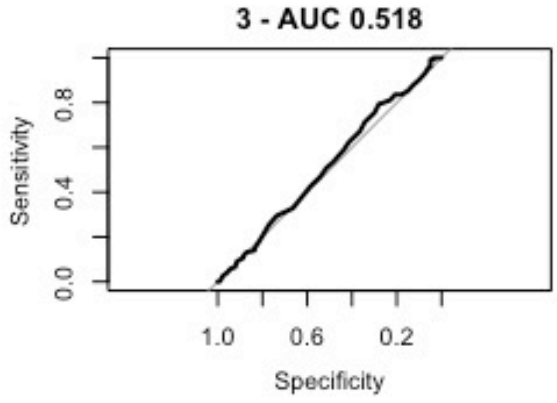
How close are we to the distribution of the data?

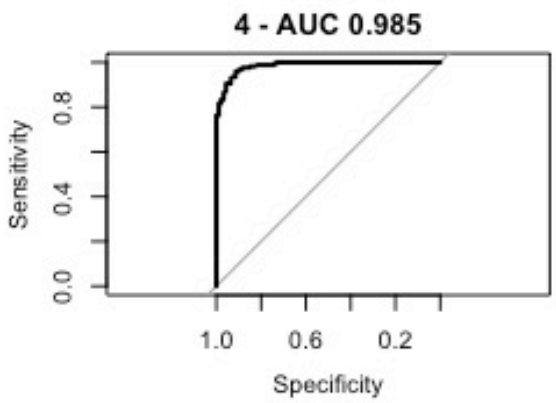


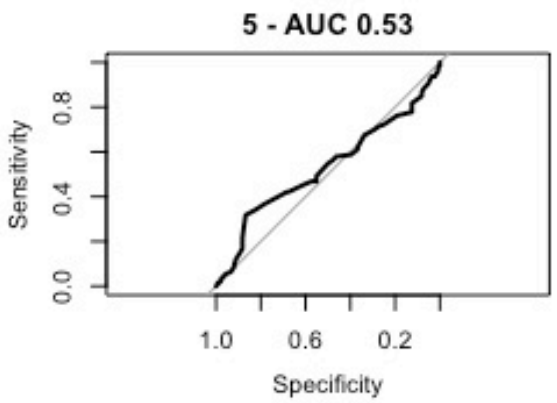
Results for Simulated Data

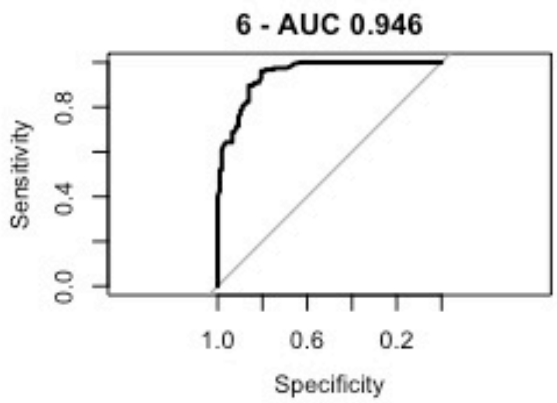




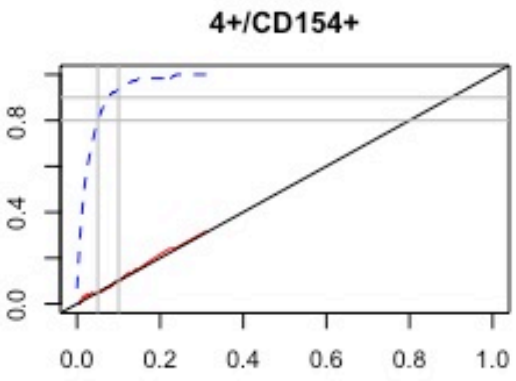


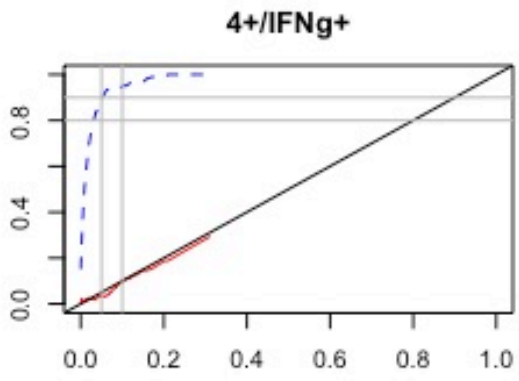


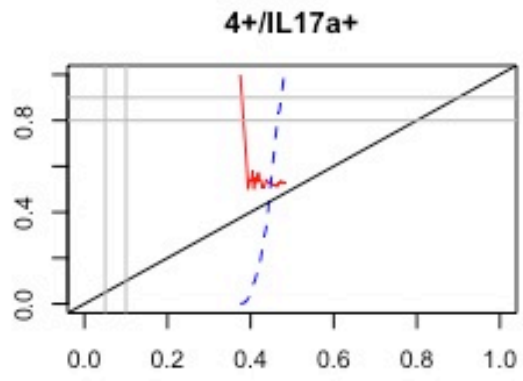


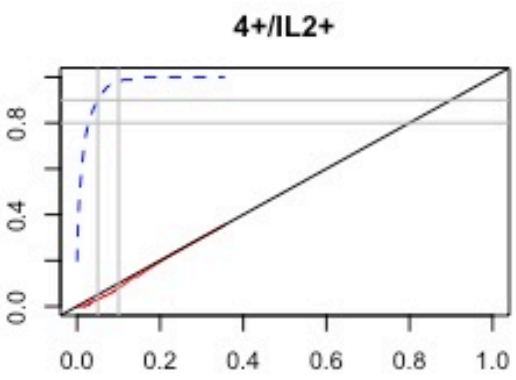


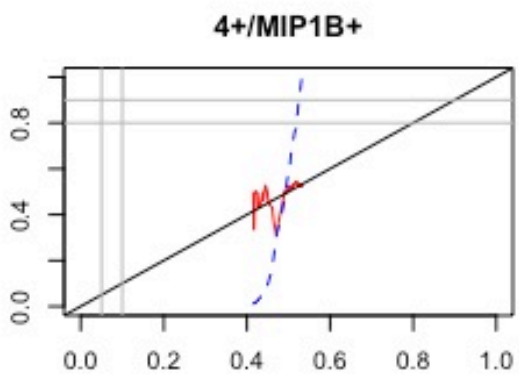
Results for Simulated Data

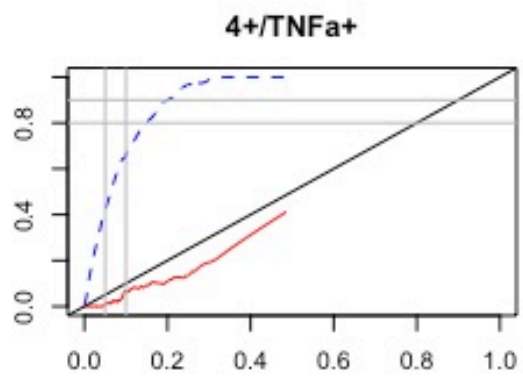












RV144 Booleans Dataset

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

RV144 Booleans Dataset

- So far we worked with marginal counts - can be obtained from bulk assays.
- single-cell measurements enable a more comprehensive understanding of cellular functions.
- **The degree of functionality** (numbered of expressed cytokines) of responsive cell-subsets has been correlated with favorable outcomes in vaccine studies.

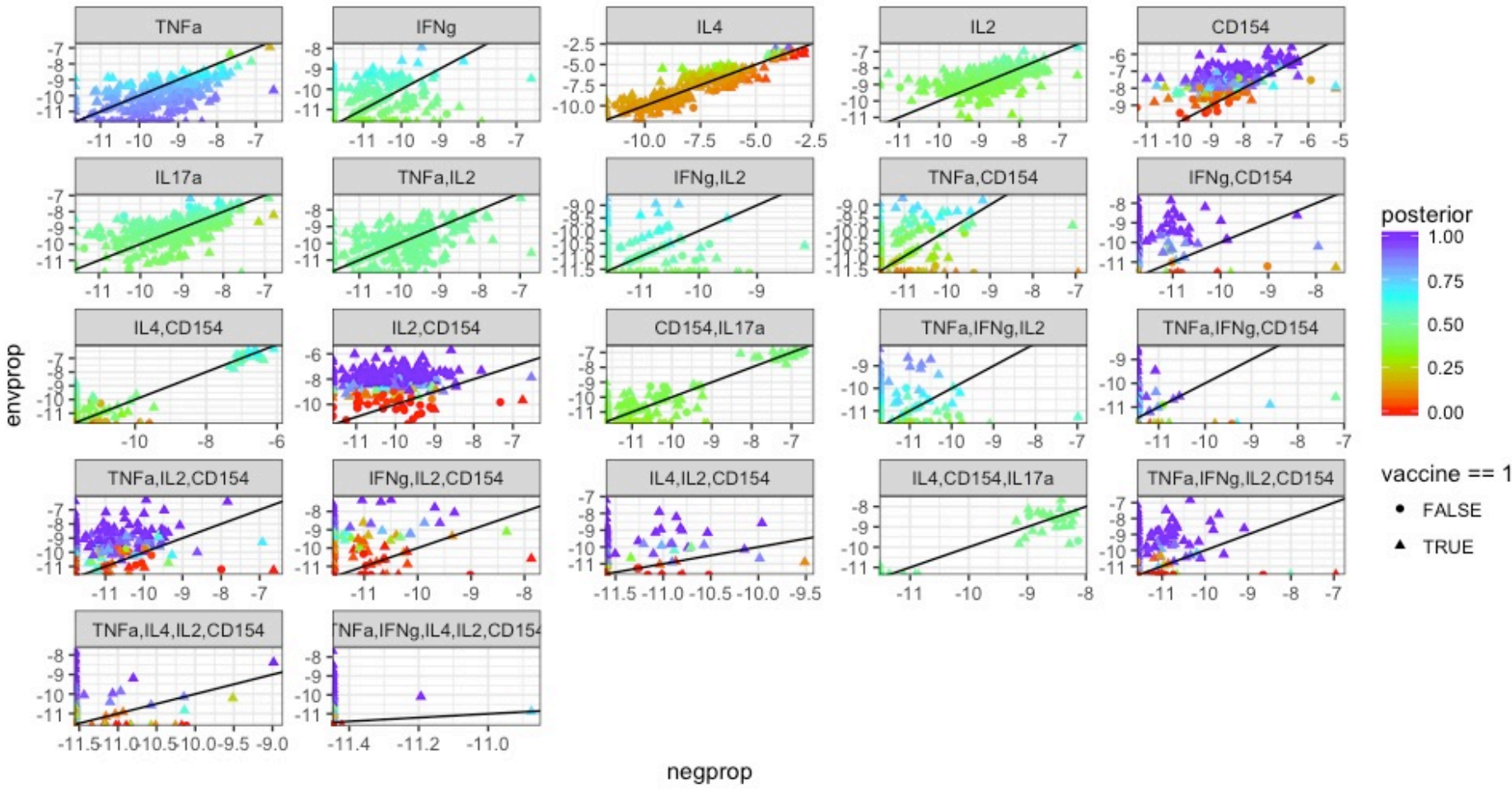
RV144 Booleans Dataset

- So far we worked with marginal counts - can be obtained from bulk assays.
- single-cell measurements enable a more comprehensive understanding of cellular functions.
- **The degree of functionality** (numbered of expressed cytokines) of responsive cell-subsets has been correlated with favorable outcomes in vaccine studies.

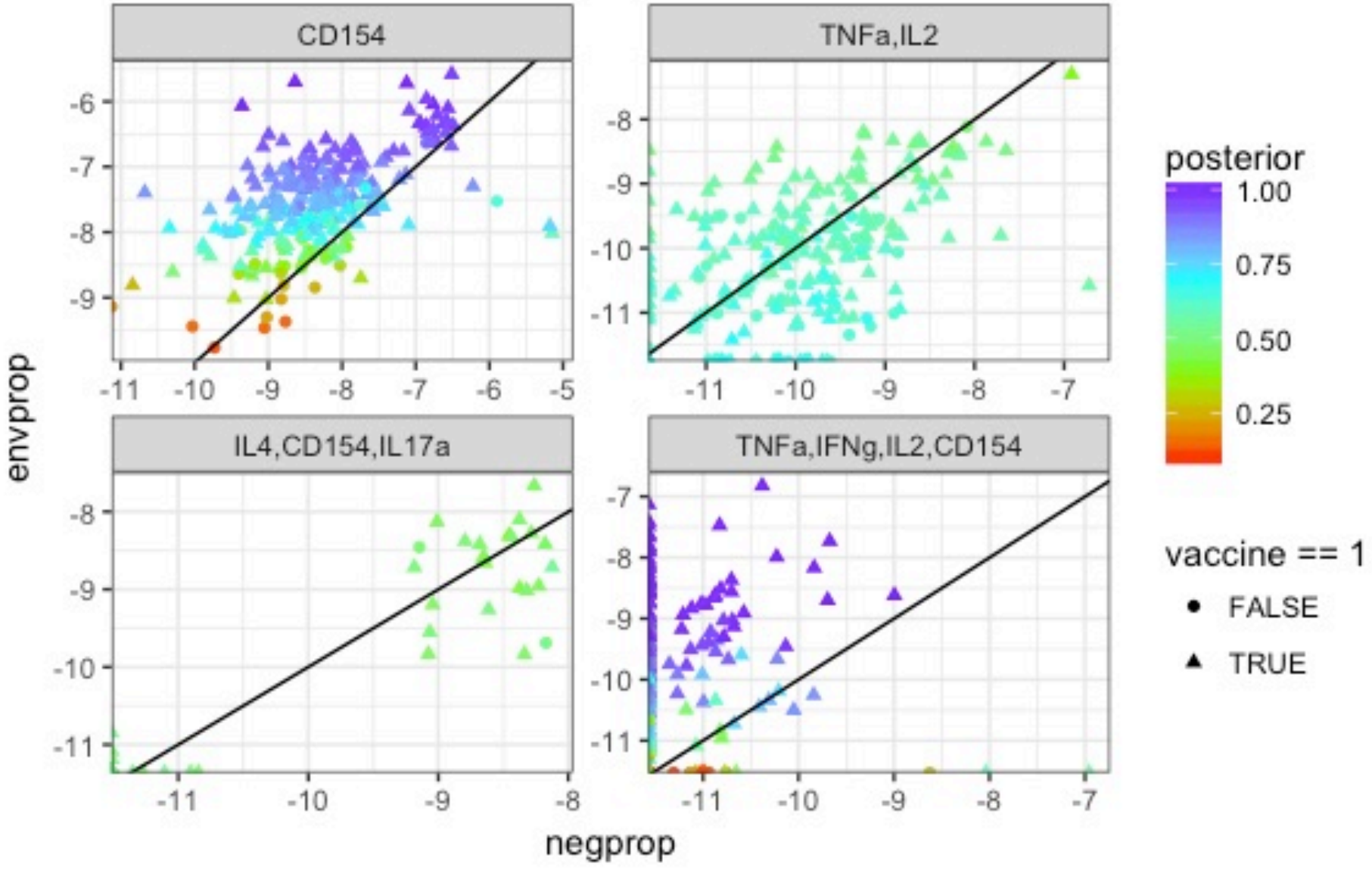
The RV144 Booleans Dataset

- **262 Subjects**
  - 226 Cases
  - 36 Controls
- **2 Types of stimulus**
  - HIV antigen
  - Negative control
- **23 types of cells with non-negligible counts.**
  - (At least two instances of $count \geq 5$)

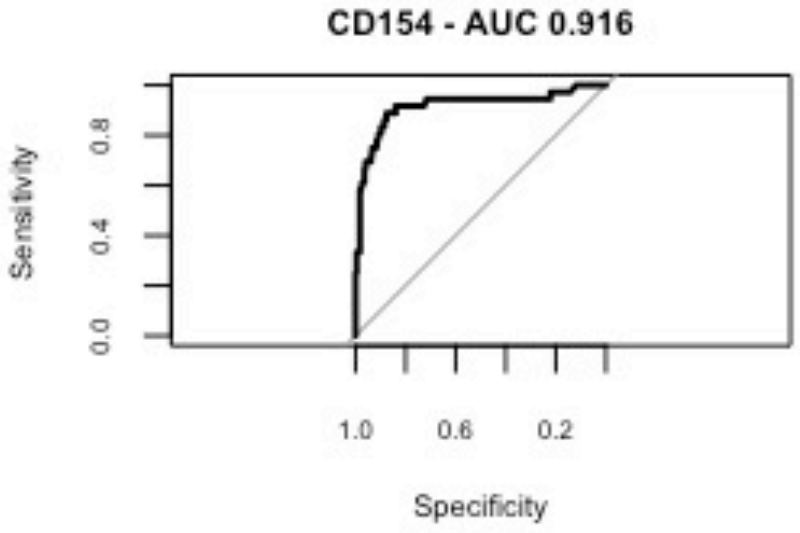
RV144 - Booleans Dataset

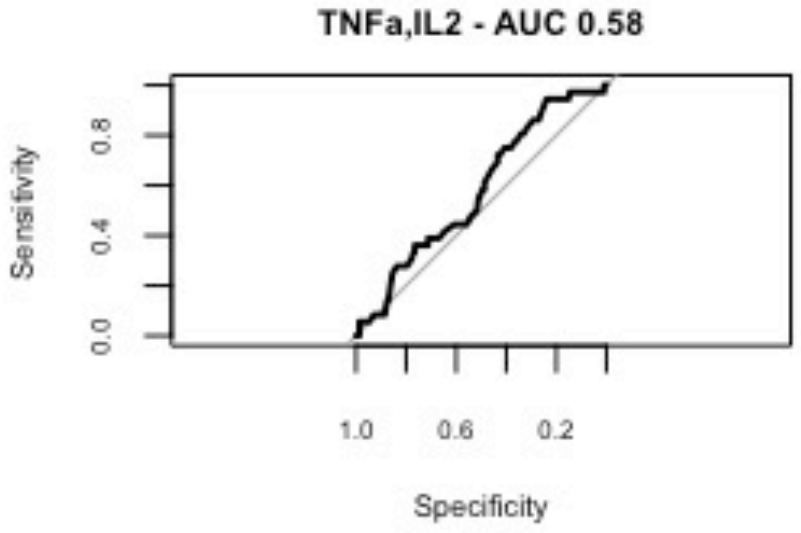


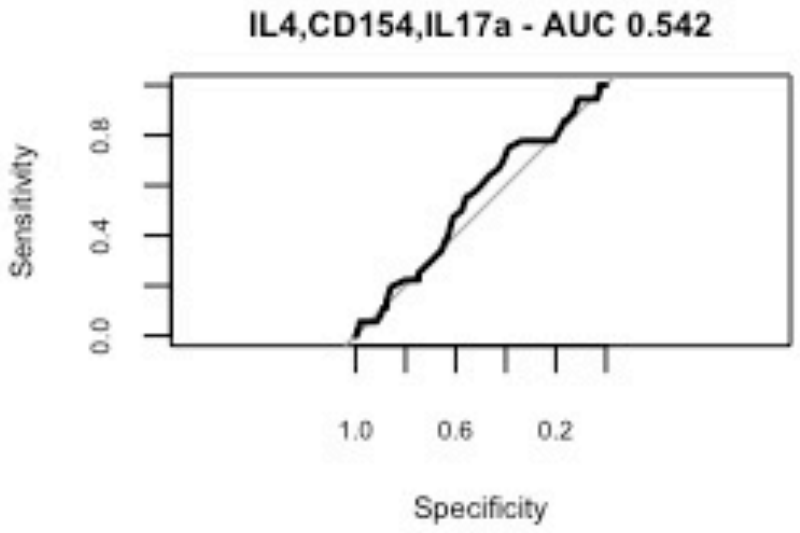
RV144 - Booleans Dataset

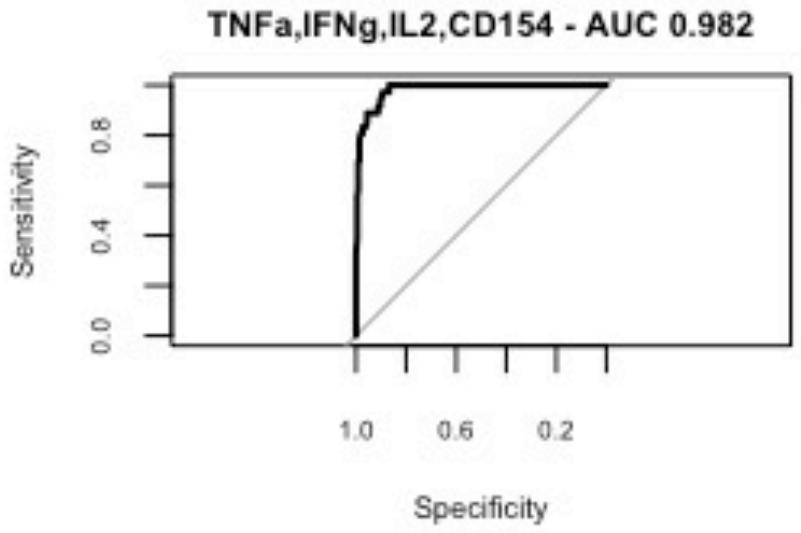


RV144 - Booleans Dataset

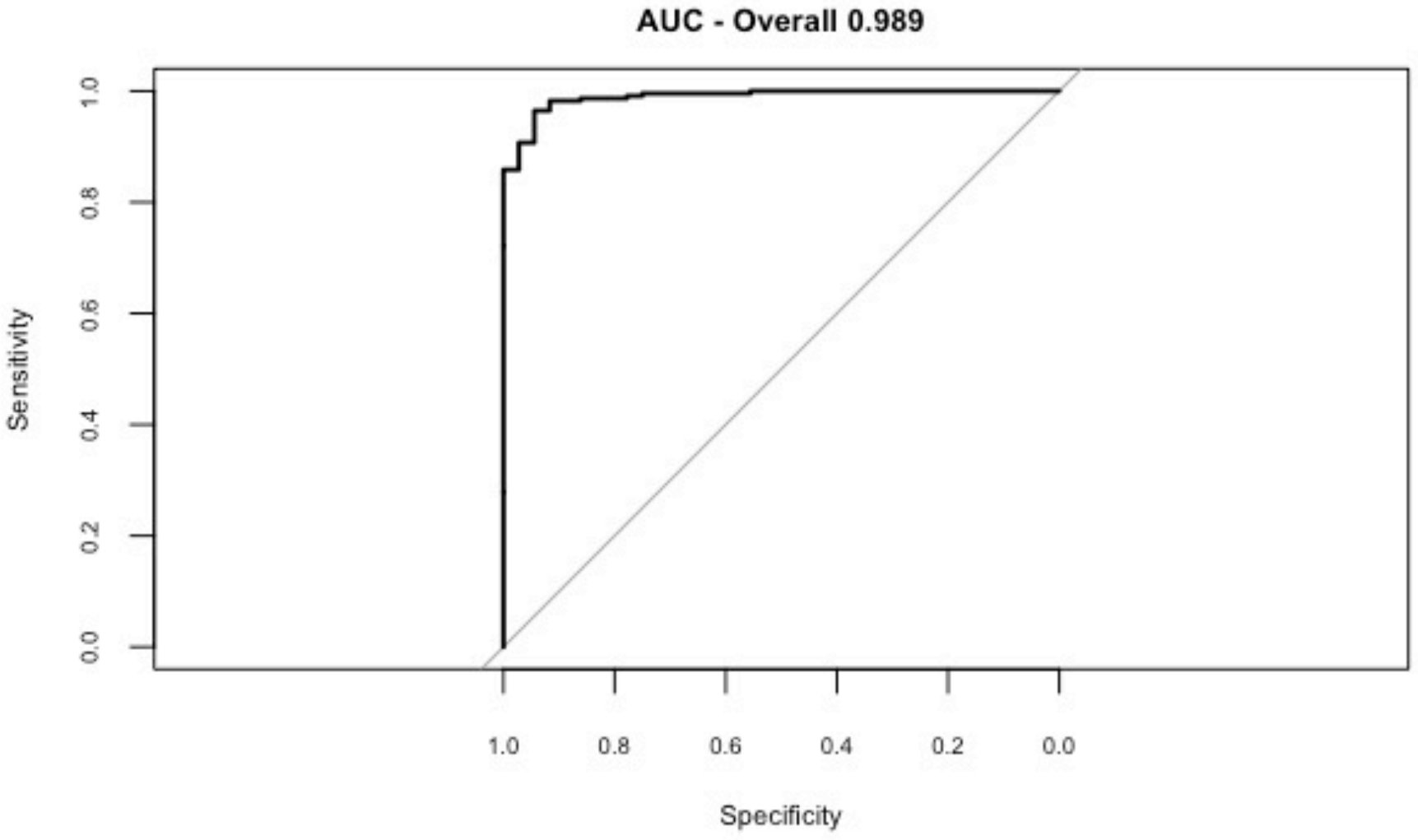








RV144 - Booleans Dataset



RV144 - Booleans Dataset

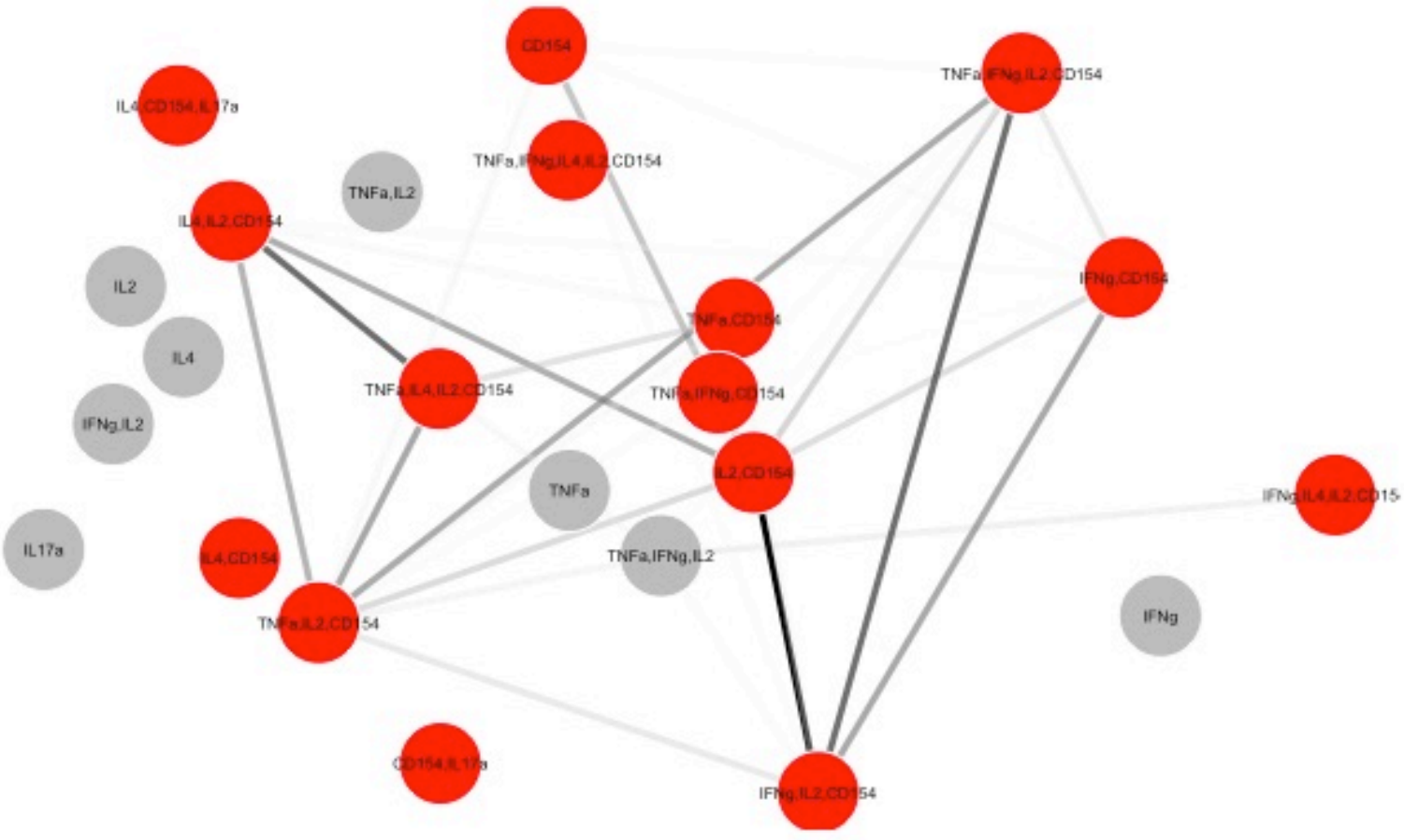


Figure: Estimated Ising Model - Red marks CD154

RV144 - Booleans Dataset

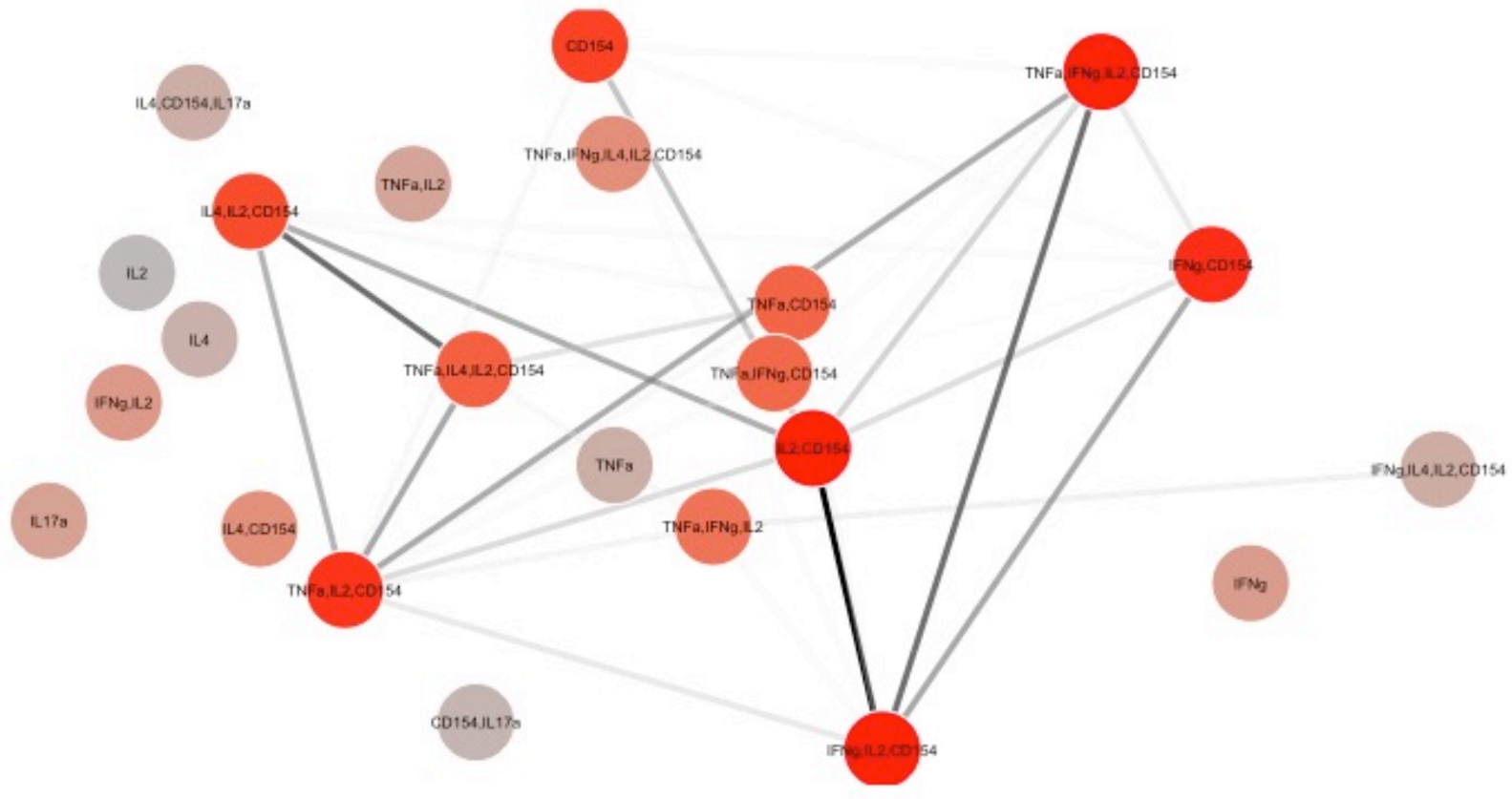
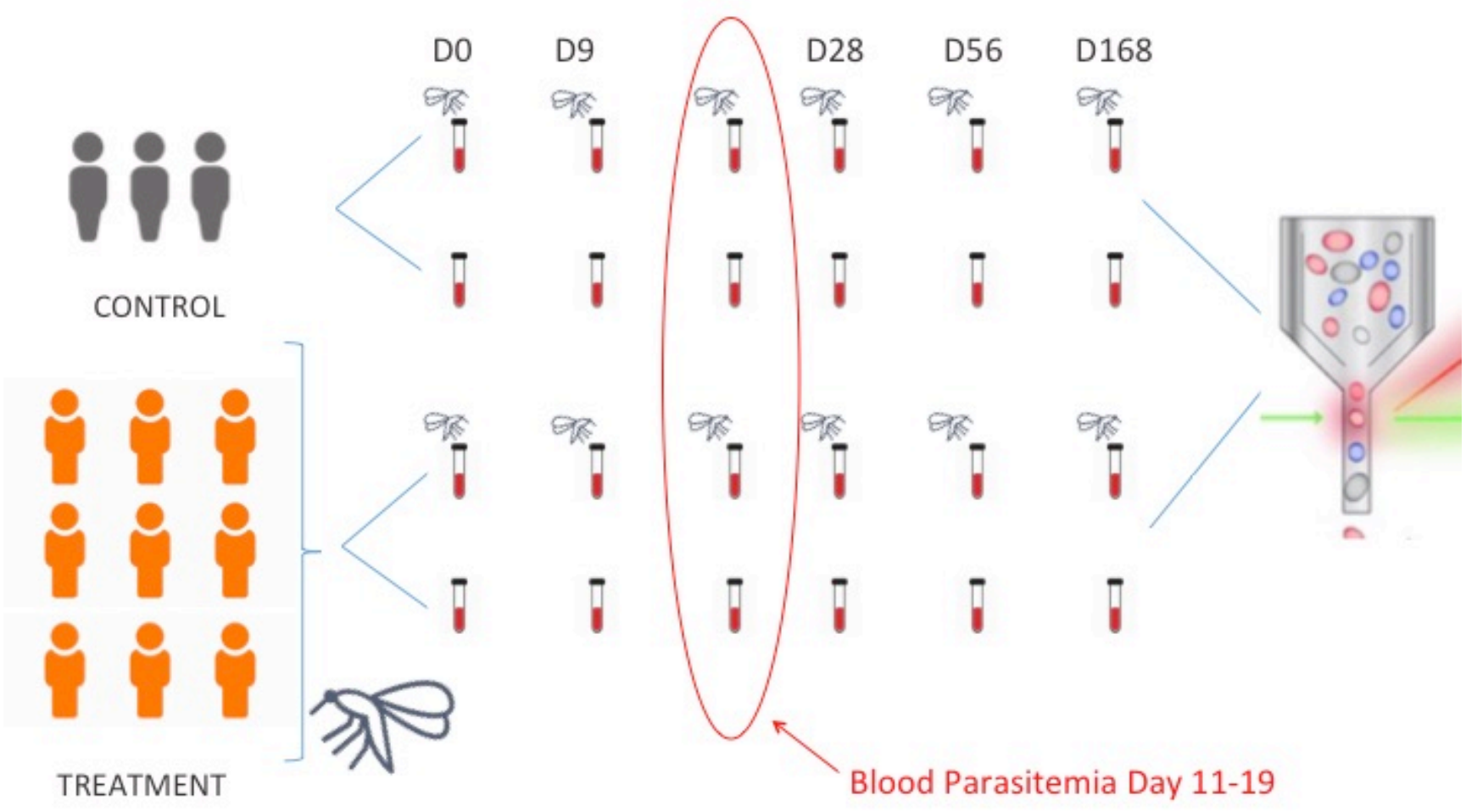


Figure: Estimated Ising Model - Red marks AUC

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

malaraFourPlusSmoothed

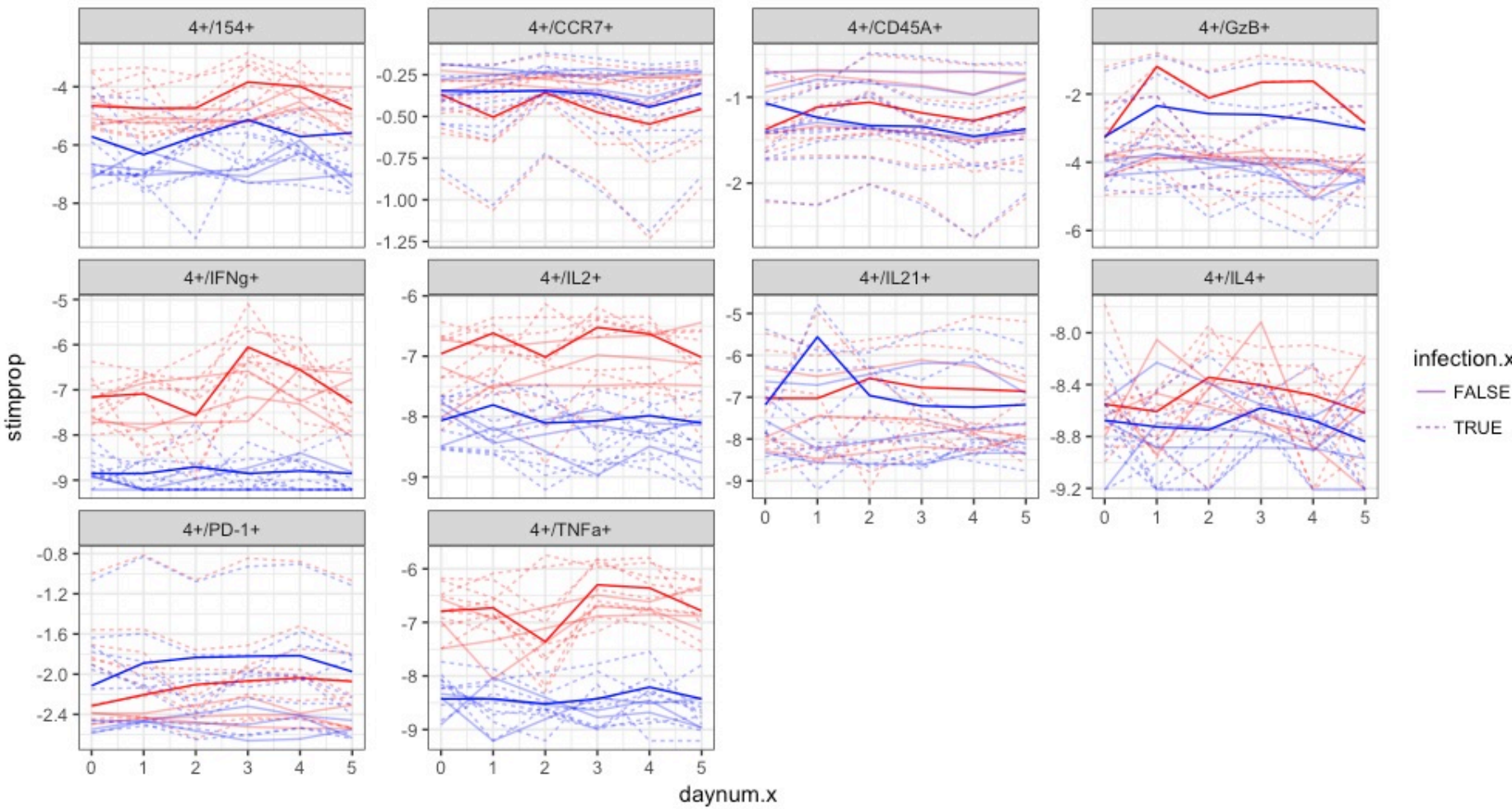


Figure: CD4 Helper Cells

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

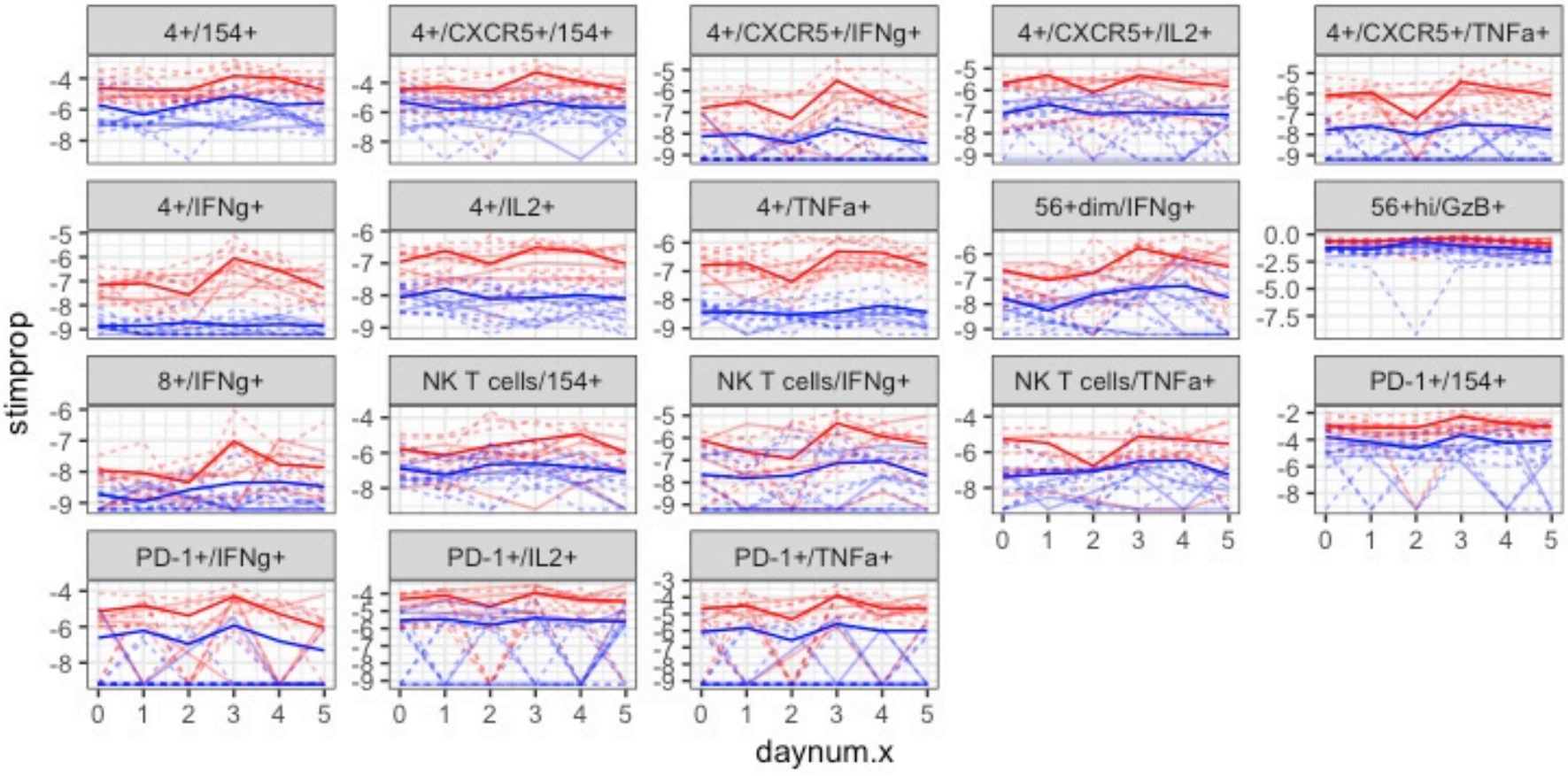


Figure: Significant Subsets

Controlled Human Malaria Infection Study

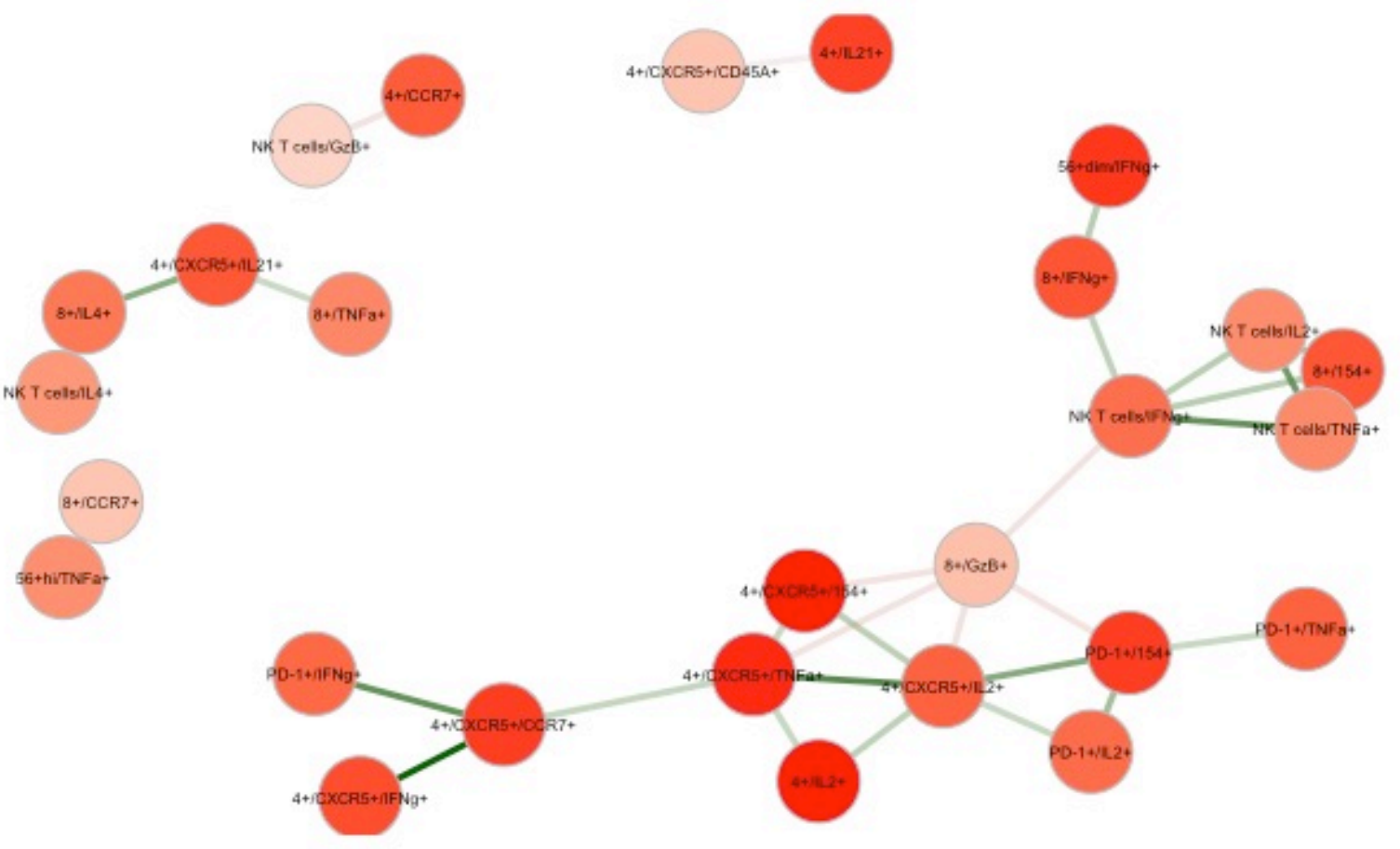


Figure: Estimated Graph - Red marks probability of response

Computation - A Difficult Likelihood

With cluster assignments and random effects known:

$$f(y_i|\nu_i, k_i) = \prod_j \prod_t f(y_{ijt}|\nu_{ij}, k_{ij})$$

The log-likelihood of the data is given by:

$$\ell(\beta, \tau, \theta, \Sigma) = \sum_{i=1}^{n} \log \left(\sum_{k \in \{0,1\}^J} P_\theta(k) \int_{\mathbb{R}^J} f_{\beta,\tau}(y_i|\nu_i, k_i) \varphi(\nu_i; 0, \Sigma) d\nu_i \right)$$

The log-likelihood is intractable for large J.

Computation - A Difficult Likelihood

With cluster assignments and random effects known:

$$f(y_i|\nu_i, k_i) = \prod_j \prod_t f(y_{ijt}|\nu_{ij}, k_{ij})$$

The log-likelihood of the data is given by:

$$\ell(\beta, \tau, \theta, \Sigma) = \sum_{i=1}^{n} \log \left(\sum_{k \in \{0,1\}^J} P_\theta(k) \int_{\mathbb{R}^J} f_{\beta,\tau}(y_i|\nu_i, k_i) \varphi(\nu_i; 0, \Sigma) d\nu_i \right)$$

The log-likelihood is intractable for large J.

Computation - How About EM?

The integrals are replaced with conditional expectations:

$$\sum_{i=1}^{n} \log \left(\sum_{k \in \{0,1\}^J} P_\theta(k|y_i) \int_{\mathbb{R}^J} f_{\beta,\tau}(y_i|\nu_i, k_i) f_\Sigma(\nu_i|y_i, k) d\nu_i \right)$$

We can use sampling to approximate the intractable integrals:

$$k_{i1}^*, ..., k_{iM}^* \sim P_\theta(k|y_i)$$

$$\nu_{i1}^* \sim f(\nu_i|y_i, k_{i1}), ..., \nu_{iM}^* \sim f(\nu_i|y_i, k_{iM})$$

$$(\beta, \tau, \theta, \Sigma) = \arg\max \sum_{i=1}^{n} \frac{1}{M} \sum_{m=1}^{M} \sum_{j} \sum_{t} \log f(y_{ijt}|\nu_{im}^*, k_{im}^*)$$

Computation - How About EM?

The integrals are replaced with conditional expectations:

$$\sum_{i=1}^{n} \log \left(\sum_{k \in \{0,1\}^J} P_\theta(k|y_i) \int_{\mathbb{R}^J} f_{\beta,\tau}(y_i|\nu_i, k_i) f_\Sigma(\nu_i|y_i, k) d\nu_i \right)$$

We can use sampling to approximate the intractable integrals:

$$k_{i1}^*, ..., k_{iM}^* \sim P_\theta(k|y_i)$$

$$\nu_{i1}^* \sim f(\nu_i|y_i, k_{i1}), ..., \nu_{iM}^* \sim f(\nu_i|y_i, k_{iM})$$

$$(\beta, \tau, \theta, \Sigma) = \arg\max \sum_{i=1}^{n} \frac{1}{M} \sum_{m=1}^{M} \sum_{j} \sum_{t} \log f(y_{ijt}|\nu_{im}^*, k_{im}^*)$$

Computation - Stochastic EM Algorithm

We alternate between a stochastic E-step and an M-step.

S - Step

- $\boldsymbol{k}^*$ - Gibbs sampler.
- $\nu^* | \boldsymbol{k}^*$ - component-wise MH algorithm.

M - Step

- β, τ - **glm, glmnet** for sparsity or **gamlss** for BB.
- θ - Pseudo-likelihood or **isingFit** for sparsity .
- $\boldsymbol{\Sigma}$ - Option for sparse estimation via **PSDE** package.

Thank you!

Questions?

AmitMeir@uw.edu