flowReMix

(temporary name)

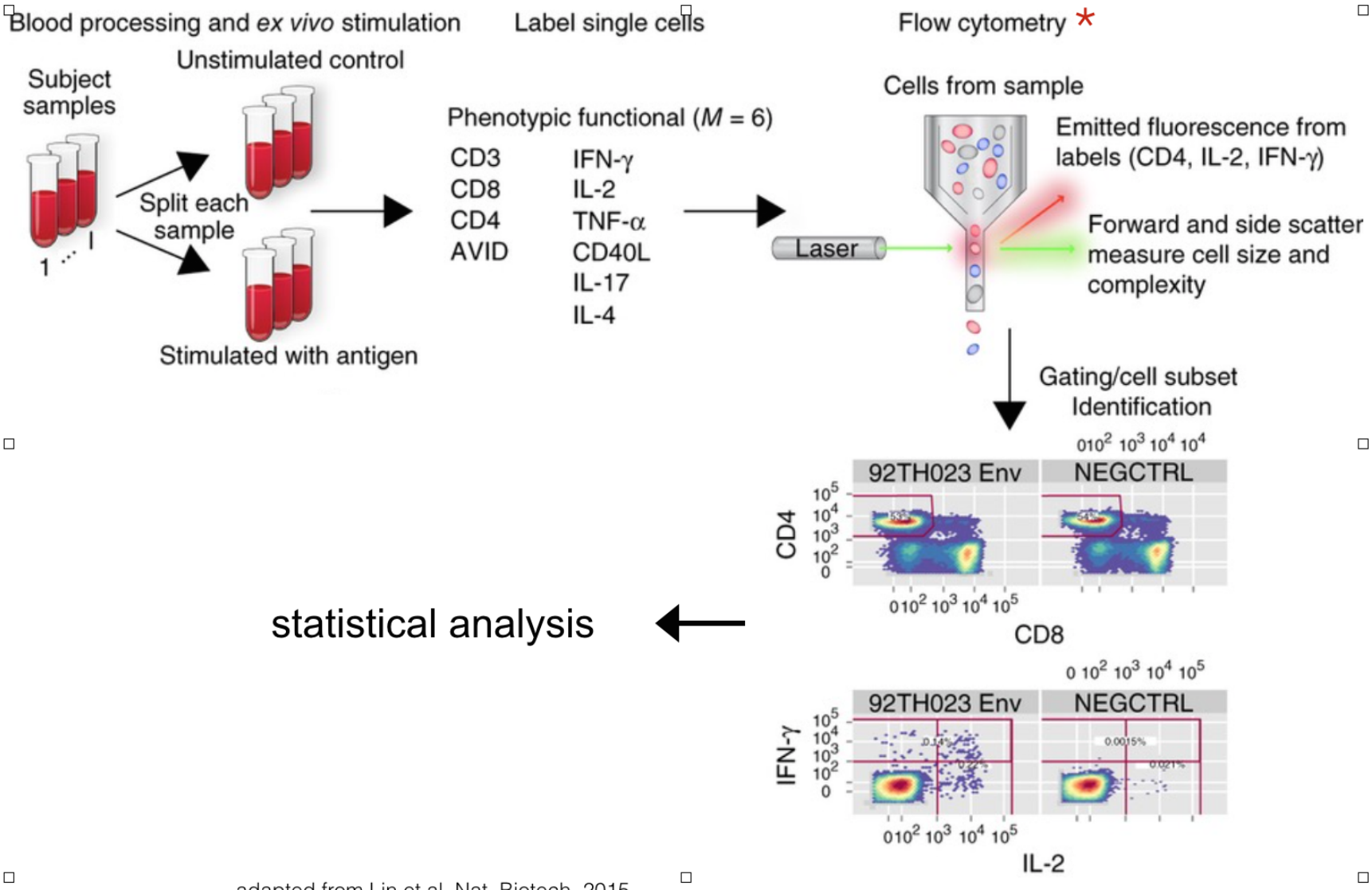
A **Mixture** of **Mixed** Beta-Binomial **Regression** Models for Analyzing **Flow**-Cytometry Count data

February 6, 2017

Outline

1. Introduction to Flow-Cytometry.
2. Challanges.
3. Motivation.
4. Models:
   - A Marginal Model.
   - Subject-Response Model.
   - Subset-Response Model.
   - Overdisperesed Subset-Response Model.
5. Data analysis.

Introduction to Cytometry Count Data



adapted from Lin et al, Nat. Biotech. 2015

The RV144 HIV Vaccine Study

- **286 Subjects**
  - 246 Cases
  - 40 Controls
- **2 Types of stimulus**
  - HIV virus
  - Negative control
- **7 types of cytokines measured**

Motivation - Unique Challanges

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

Over-Dispersion

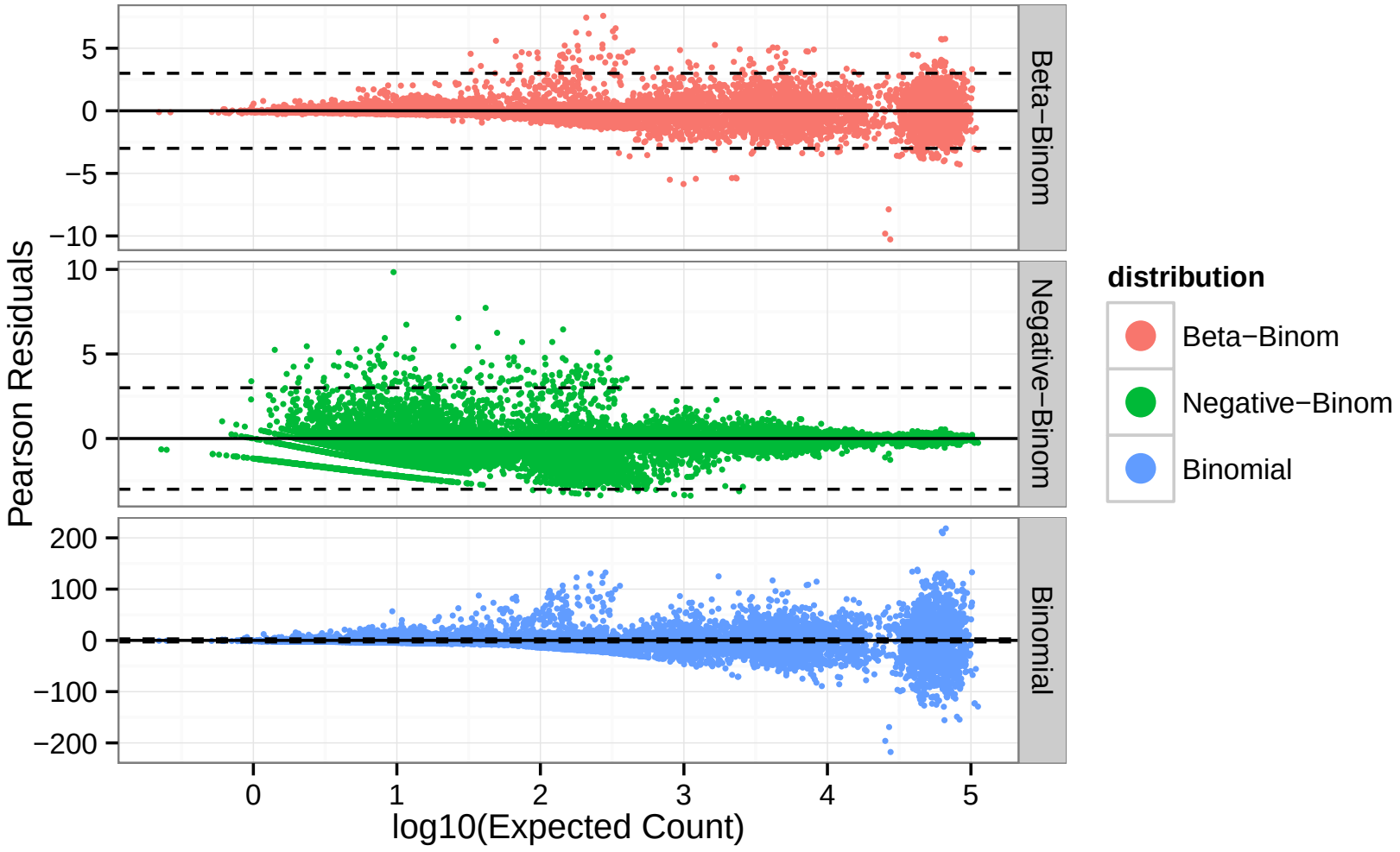


Figure: Standartized Residuals for RV144.

Dependence

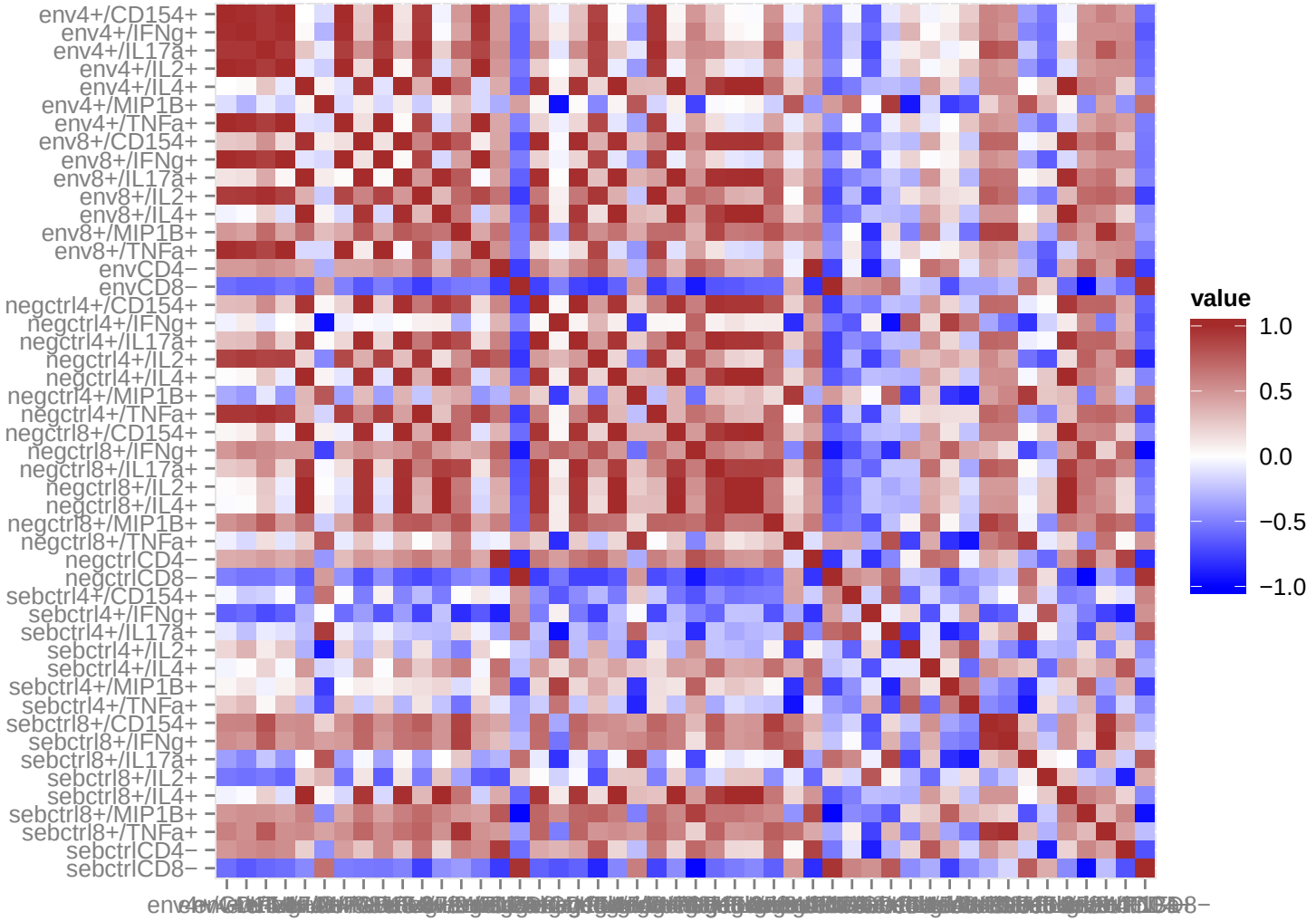


Figure: Residual Correlations for RV144.

Dependence

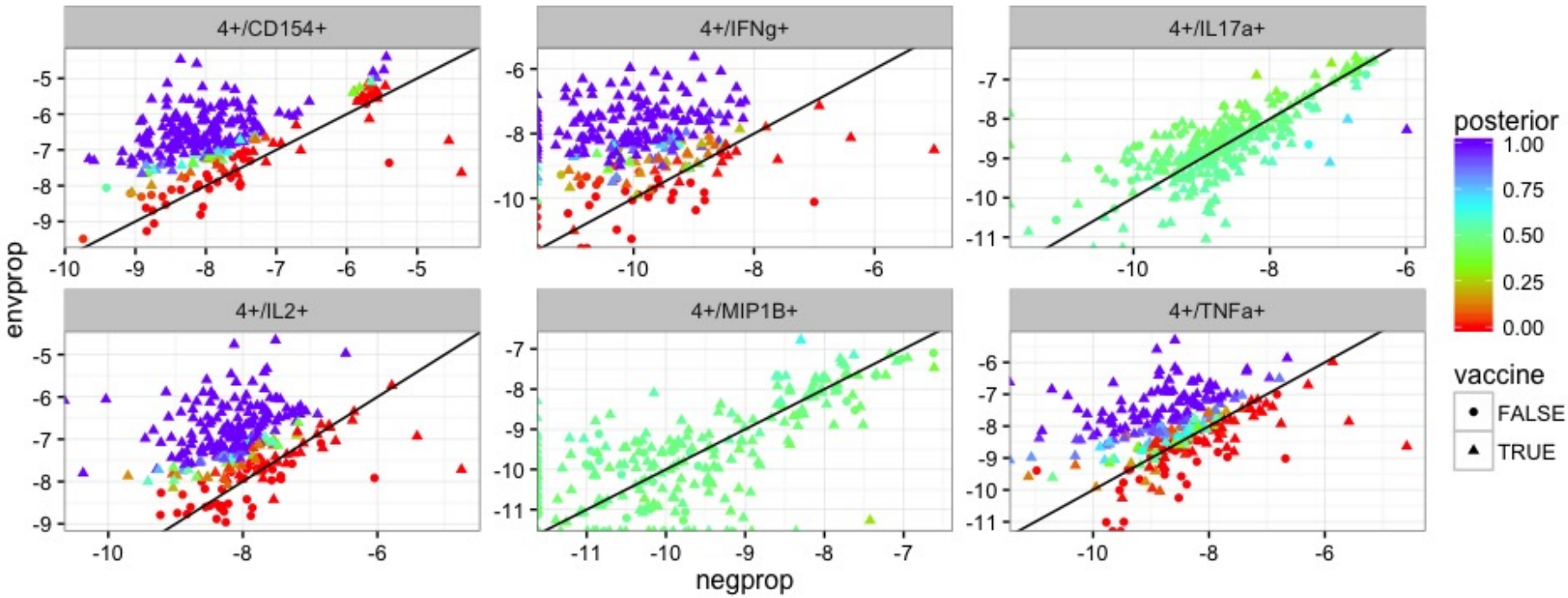


Figure: Scatter Plots

Motivation - A Regression Model

- Covariates unrelated to the treatment.
- Batch effects.
- Longitudal data.
- Joint modeling of several cell-subsets.
- Modeling Dependence.

A Marginal Model - Single Subset

Indexing: **i**-subject, **l**- stimulation.

- Binomial count data $\Rightarrow$ Logistic model.

$$\text{logit}(p_{il}) = X_{il}\beta$$

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

- Dependence $\Rightarrow$ 'random' subject baseline:

$$\text{logit}(p_{il}) = X_{il}\beta + \nu_i$$

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

A Marginal Model - Single Subset

Indexing: **i**-subject, **l**- stimulation.

- Binomial count data $\Rightarrow$ Logistic model.

$$\text{logit}(p_{il}) = X_{il}\beta$$

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

- Dependence $\Rightarrow$ 'random' subject baseline:

$$\text{logit}(p_{il}) = X_{il}\beta + \nu_i$$

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

A Marginal Model - Single Subset

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

- Non-response $\Rightarrow$ Mixture-Model:

$$\text{logit}(p_{ilk}) = X_{il}\beta + T_{il}\tau_k + \nu_i$$

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

Marginal Model - Estimation via Stochastic EM

The likelihood can be written as:

$$\sum_{i=1}^{n} \log \left(\sum_{k=0}^{1} \int_{\mathbb{R}} \prod_{l} f_k(y_{il}|\nu_i) \varphi(\nu_i) d\nu_i \right)$$

In a Stochastic EM we sample, $\nu_i^*, k_i^* \sim \nu_i, k_i|y_i$ and maximize:

$$\max_{\theta} \sum_{i=1}^{n} \sum_{l=1}^{l} \left(\log f_{k_i^*}(y_{il}|\nu_i^*) + \log P(k_i^*) \right)$$

Marginal Model - Estimation via Stochastic EM

The likelihood can be written as:

$$\sum_{i=1}^{n} \log \left(\sum_{k=0}^{1} \int_{\mathbb{R}} \prod_{l} f_k(y_{il}|\nu_i)\varphi(\nu_i) d\nu_i \right)$$

In a Stochastic EM we sample, $\nu_i^*, k_i^* \sim \nu_i, k_i|y_i$ and maximize:

$$\max_{\theta} \sum_{i=1}^{n} \sum_{l=1}^{l} \left(\log f_{k_i^*}(y_{il}|\nu_i^*) + \log P(k_i^*) \right)$$

Marginal Model - Estimation via Stochastic EM

How to sample ν^* and k^*:

- Integrate ν^* out to sample k^*:

$$P(k_1^* = 1) = \frac{\int \prod_I f_1(y_{il}|\nu)\varphi(\nu)d\nu}{\int \prod_I f_1(y_{il}|\nu)\varphi(\nu)d\nu + \int \prod_I f_0(y_{il}|\nu)\varphi(\nu)d\nu}$$

- Given k^* sample ν^* using Metropolis Hastings.

$$r^t \sim N(\nu^{t-1}, \sigma^2),$$

$$P(\nu^t = r^t) = \frac{\prod_I f_{k_i^*}(y_{il}|r^t)\varphi(r^t)}{\prod_I f_{k_i^*}(y_{il}|\nu^{t-1})\varphi(\nu^{t-1})}$$

Given ν^* and k^* the maximization of the likelihood is as simple as fitting a logistic regression.

Marginal Model - Estimation via Stochastic EM

How to sample ν^* and k^*:

- Integrate ν^* out to sample k^*:

$$P(k_1^* = 1) = \frac{\int \prod_I f_1(y_{iI}|\nu)\varphi(\nu)d\nu}{\int \prod_I f_1(y_{iI}|\nu)\varphi(\nu)d\nu + \int \prod_I f_0(y_{iI}|\nu)\varphi(\nu)d\nu}$$

- Given k^* sample ν^* using Metropolis Hastings.

$$r^t \sim N(\nu^{t-1}, \sigma^2),$$

$$P(\nu^t = r^t) = \frac{\prod_I f_{k_i^*}(y_{iI}|r^t)\varphi(r^t)}{\prod_I f_{k_i^*}(y_{iI}|\nu^{t-1})\varphi(\nu^{t-1})}$$

Given ν^* and k^* the maximization of the likelihood is as simple as fitting a logistic regression.

Marginal Model - Estimation via Stochastic EM

How to sample ν^* and k^*:

- Integrate ν^* out to sample k^*:

$$P(k_1^* = 1) = \frac{\int \prod_l f_1(y_{il}|\nu)\varphi(\nu)d\nu}{\int \prod_l f_1(y_{il}|\nu)\varphi(\nu)d\nu + \int \prod_l f_0(y_{il}|\nu)\varphi(\nu)d\nu}$$

- Given k^* sample ν^* using Metropolis Hastings.

$$r^t \sim N(\nu^{t-1}, \sigma^2),$$

$$P(\nu^t = r^t) = \frac{\prod_l f_{k_i^*}(y_{il}|r^t)\varphi(r^t)}{\prod_l f_{k_i^*}(y_{il}|\nu^{t-1})\varphi(\nu^{t-1})}$$

Given ν^* and k^* the maximization of the likelihood is as simple as fitting a logistic regression.

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

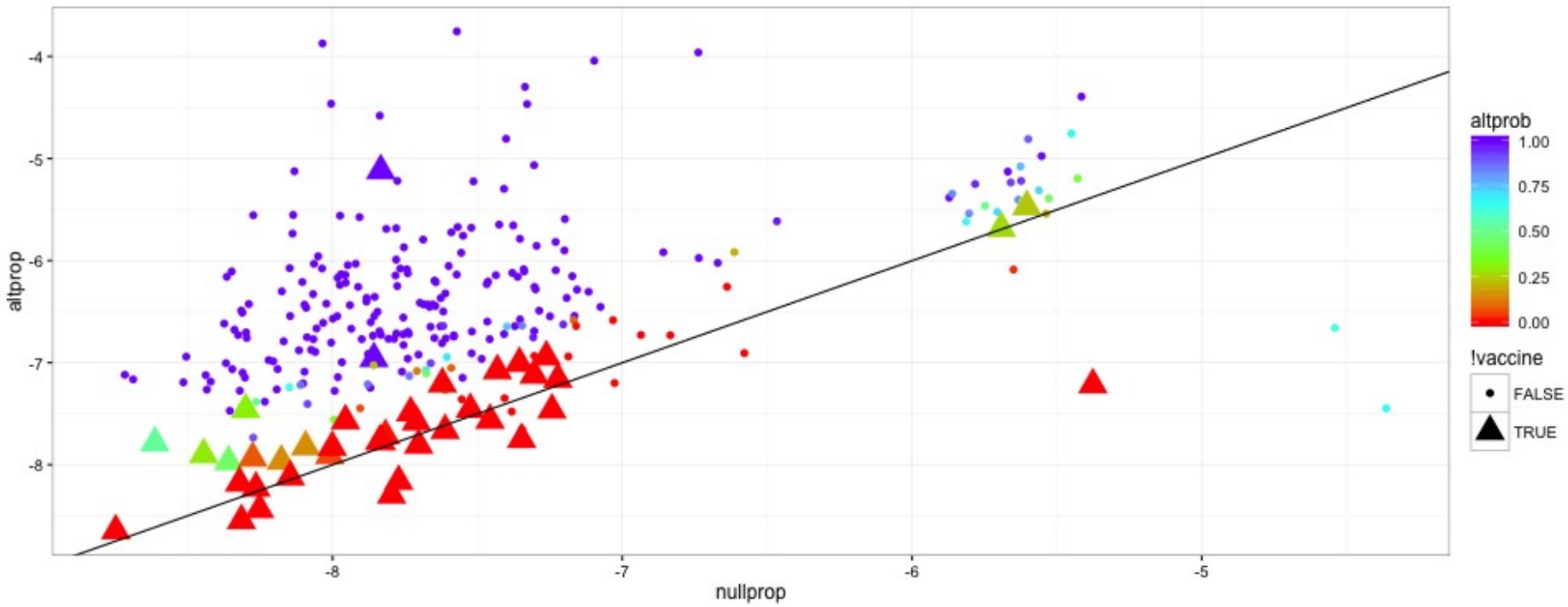


Figure: Scatter plot for T4+/CD154+ - Marginal Model

Marginal Model - Results

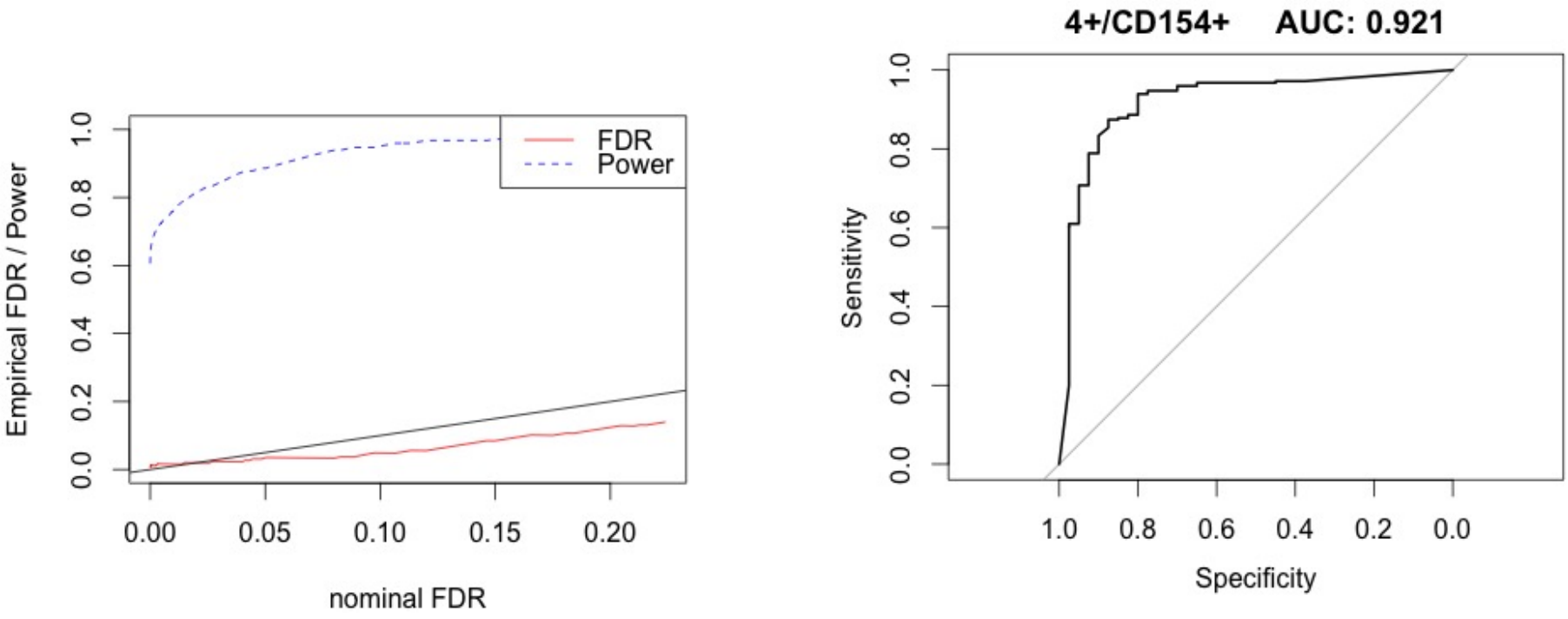


Figure: ROC/FDR plots for T4+/CD154+ - Marginal Model

Comparison with MIMOSA

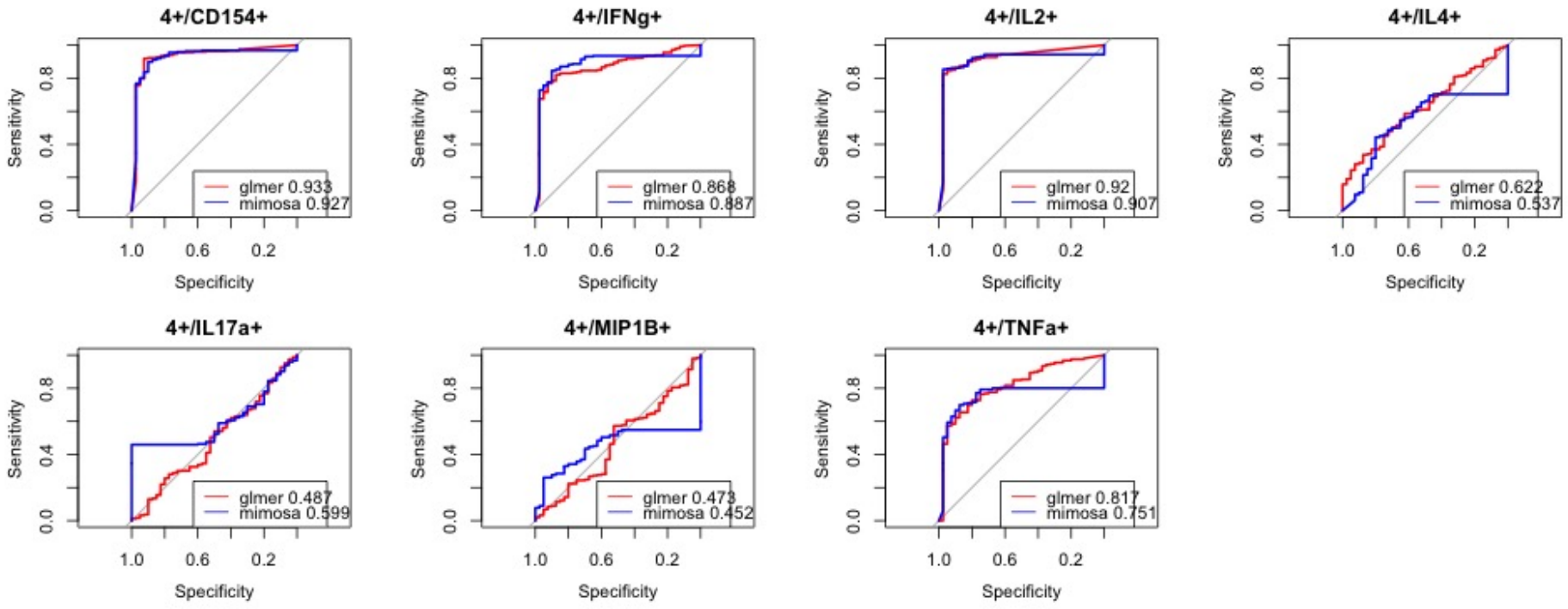


Figure: Comparison with Mimosa

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.
- 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.
- We might be able to improve classification of response by looking at several cell-subsets at once.

Subject-Response Model

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

$$\text{logit}(p_{ijlk}) = X_{ijl}\beta + T_{ijl}\tau_{k_i} + \nu_{ij},$$

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

$$y_{ijlk} \sim \text{Binom}(N_{il}, p_{ijlk}).$$

Model is same as the marginal model except that the random effect is multivariate.

Subject-Response Model

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

$$\text{logit}(p_{ijlk}) = X_{ijl}\beta + T_{ijl}\tau_{k_i} + \nu_{ij},$$

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

$$y_{ijlk} \sim \text{Binom}(N_{il}, p_{ijlk}).$$

Model is same as the marginal model except that the random effect is multivariate.

Subject-Response Model - Results

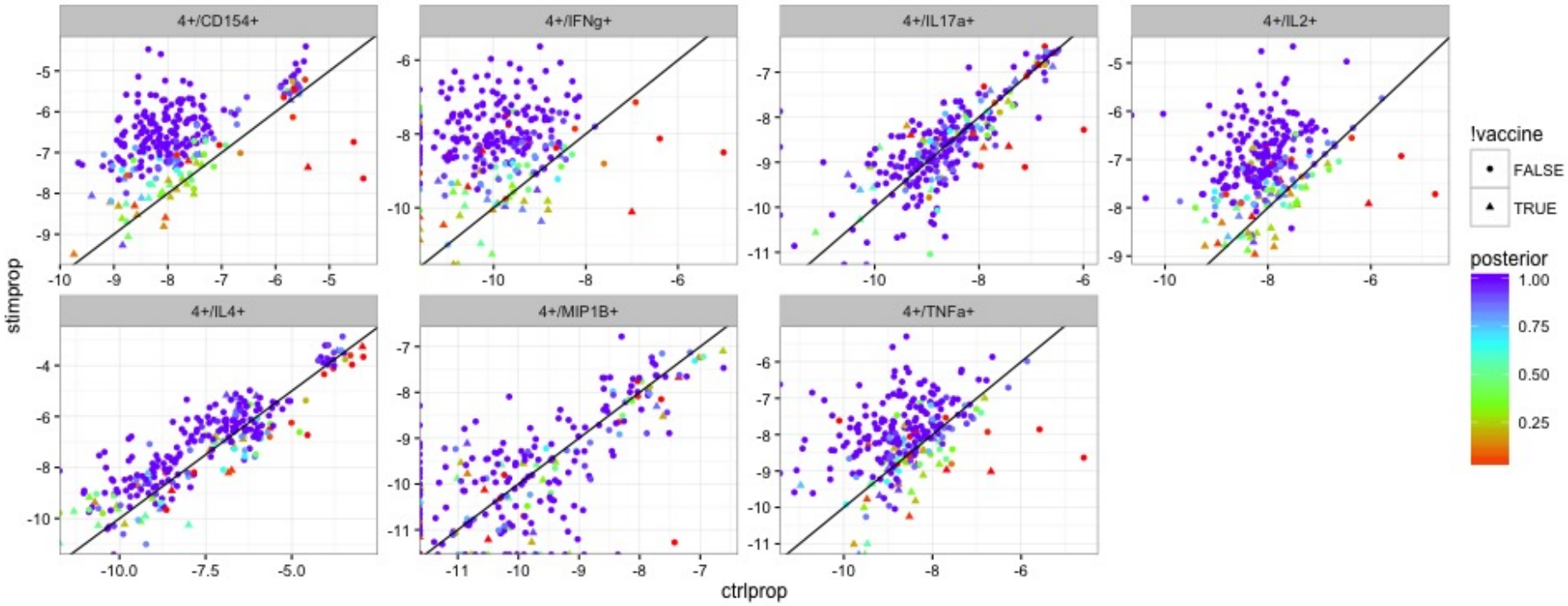
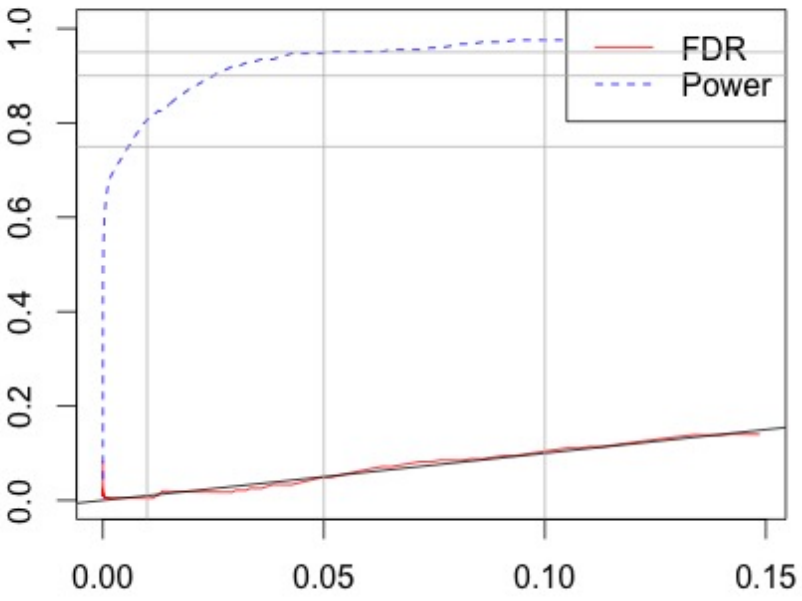


Figure: Scatter plot for Subject-Response Model

Subject-Response Model - Results



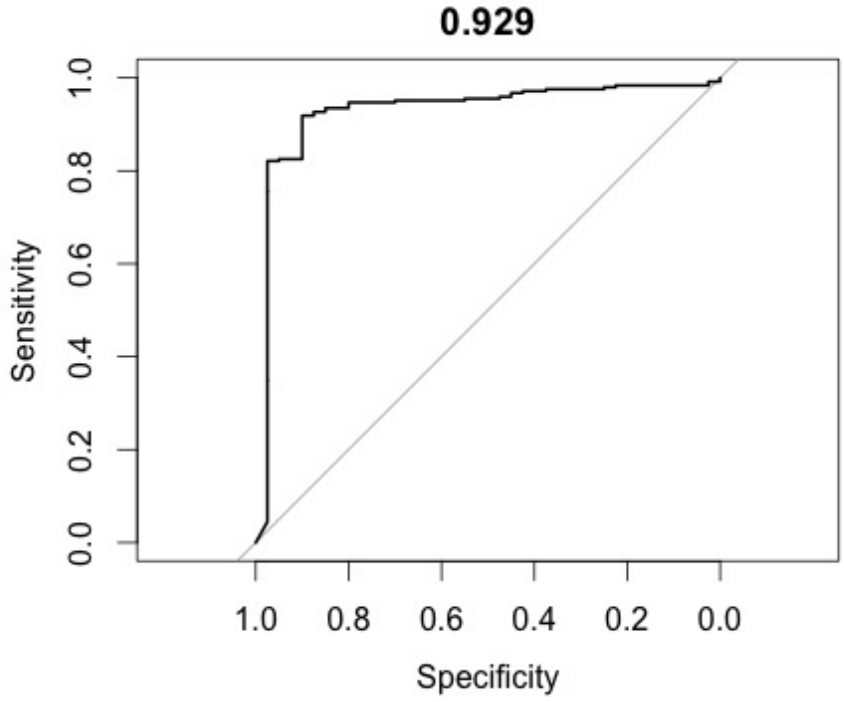


Figure: ROC for Subject-Response Model

Subset-Response Model - Motivation

- The main scientific interest is in the cell-subset level.
- Some cell-subsets exhibit very strong response and other none at all.
- Response rates may vary across subsets.
- The identity of responding cell-subsets may better predict successful immunization.

Subset-Response Model

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

Cluster assignment is defined by a $\{0,1\}^p$ vector with 1 indicating a response 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.

Subset-Response Model

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

Cluster assignment is defined by a $\{0,1\}^p$ vector with 1 indicating a response 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.

Subset-Response Model

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

Cluster assignment is defined by a $\{0,1\}^p$ vector with 1 indicating a response 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.

Subset-Response Model - Results

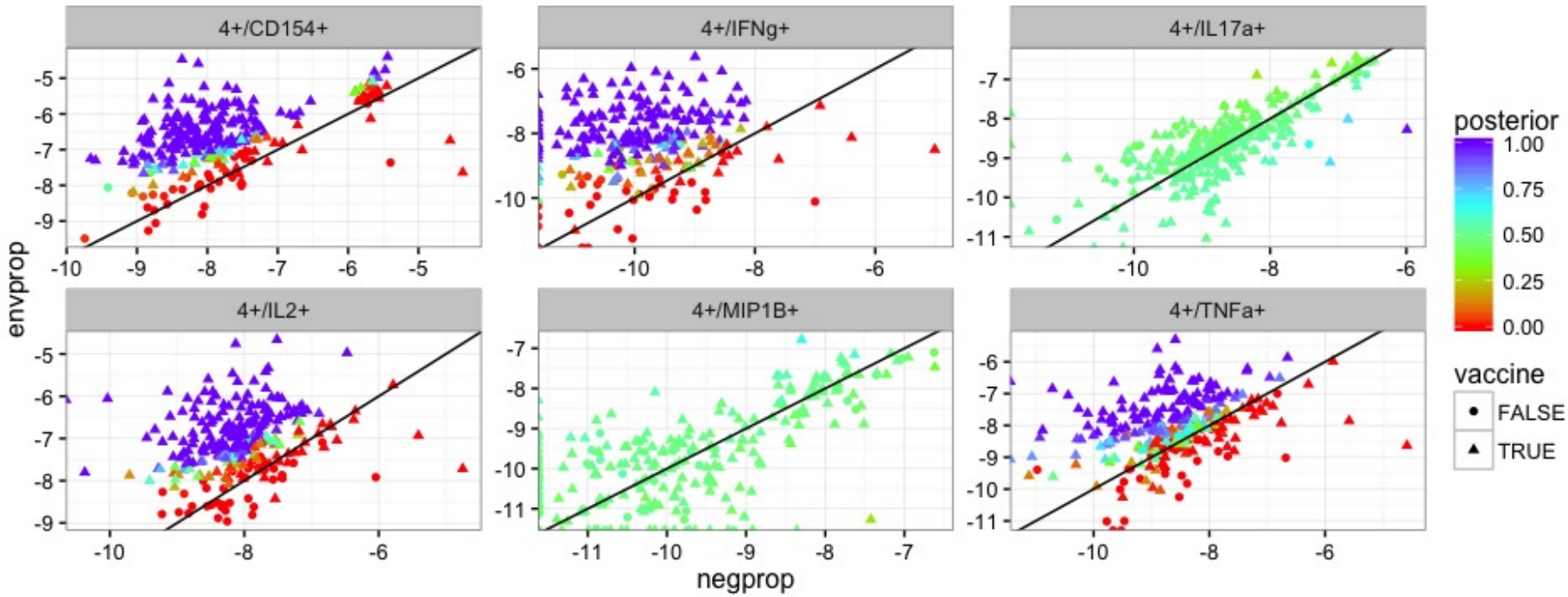


Figure: Scatter Plot for Subset-Response Model

Subset-Response Model - Results

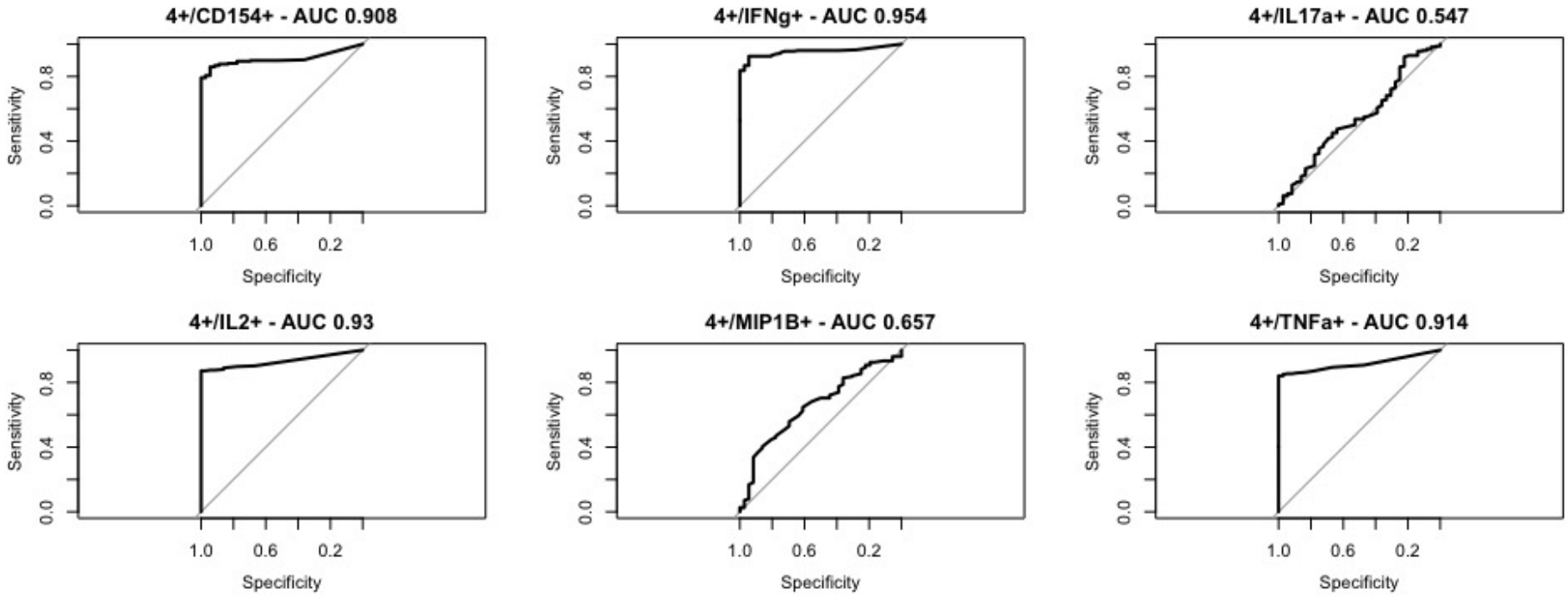


Figure: ROC for Subset-Response Model

Subset-Response Model - Results

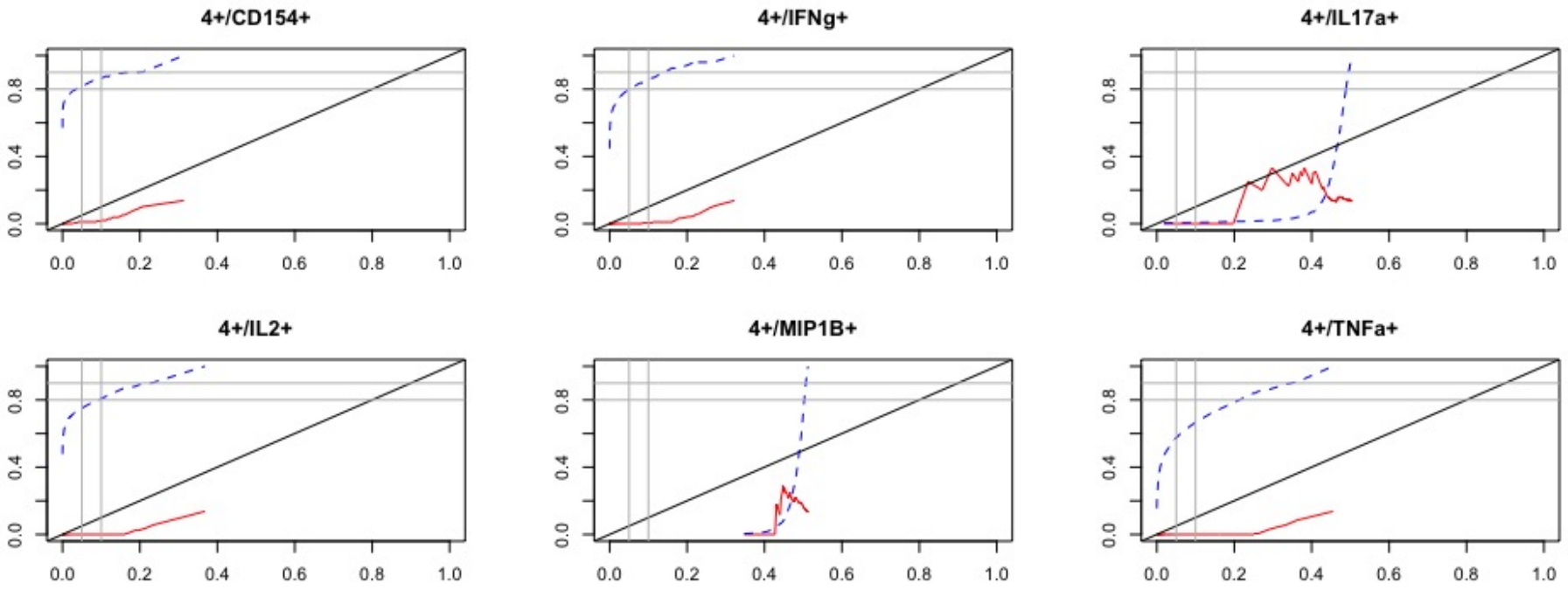


Figure: FDR for Subset-Response Model

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?

Simulated Binomial Data - Results

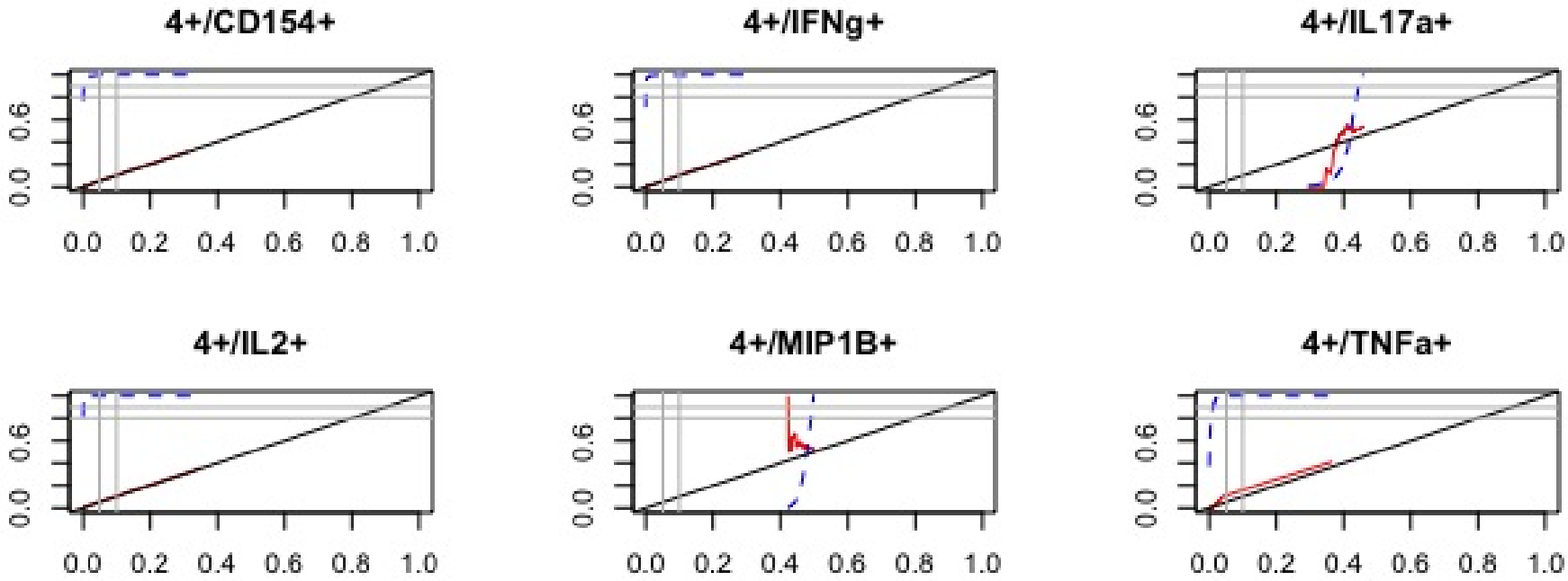


Figure: FDR for Simulated Binomial Data

Simulated Binomial Data - Results

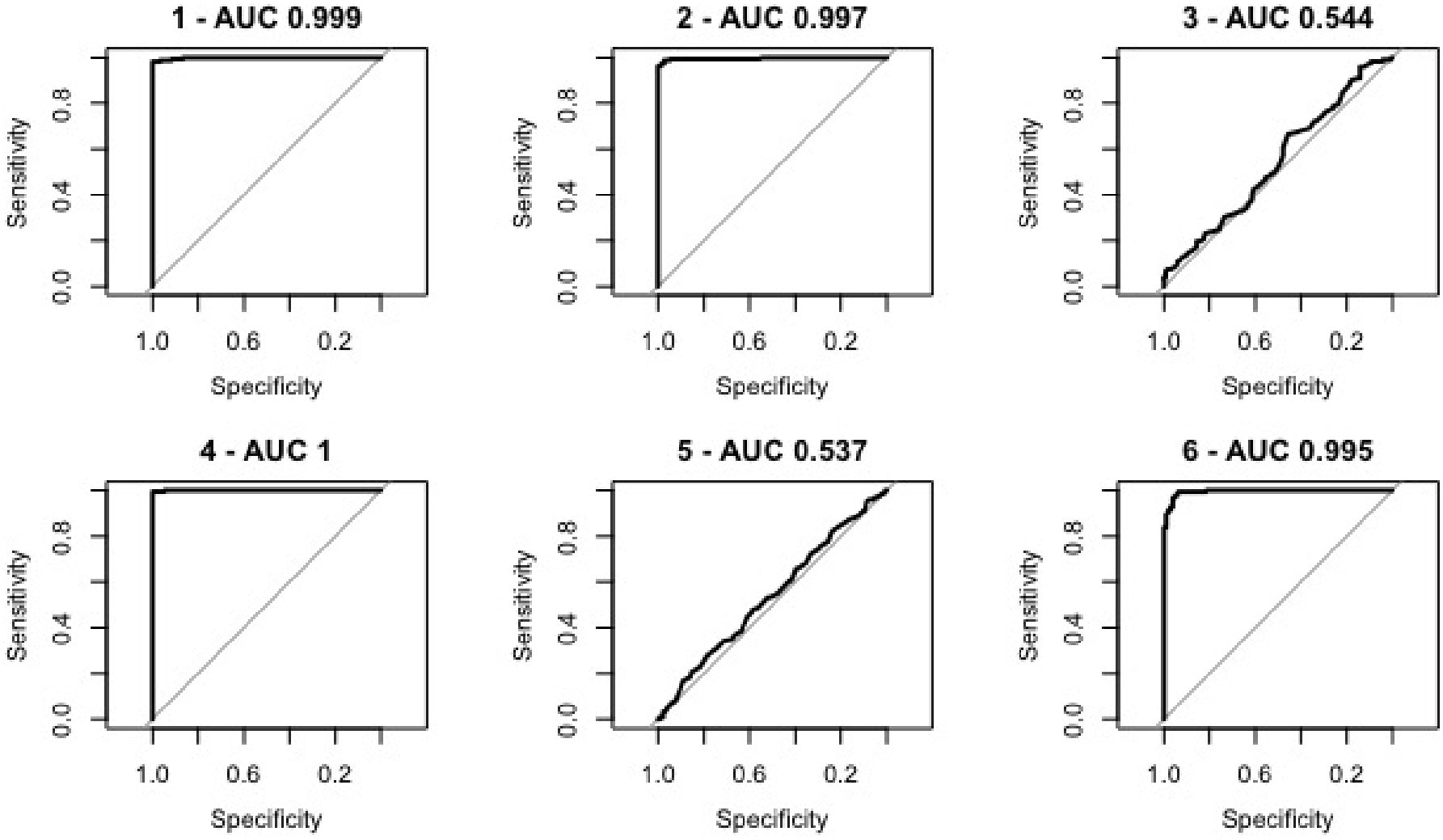


Figure: ROC for Simulated Binomial Data

Simulated Binomial Data - Results

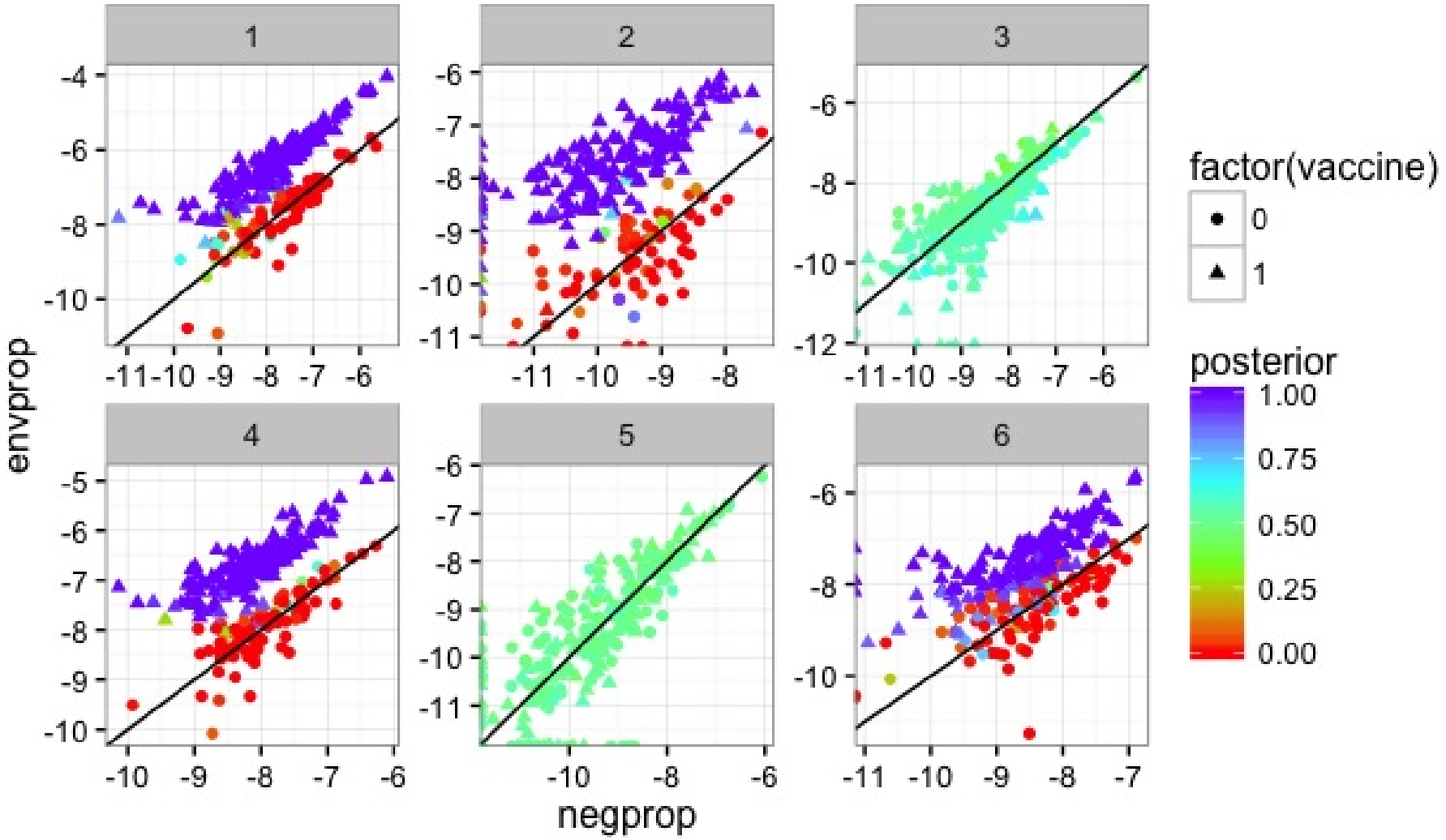


Figure: Scatter plot for Simulated Binomial Data

An Overdispersed Model

We are clearly missing some variablility...

Assume a Beta-Binomial Model:

$$\text{logit}(\mu) = X\beta + T\tau + \nu,$$

$$p \sim \text{Beta}(M\mu, M(1-\mu)), \qquad M > 0,$$

$$y \sim \text{Binom}(N, p).$$

An Overdispersed Model - Recap

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

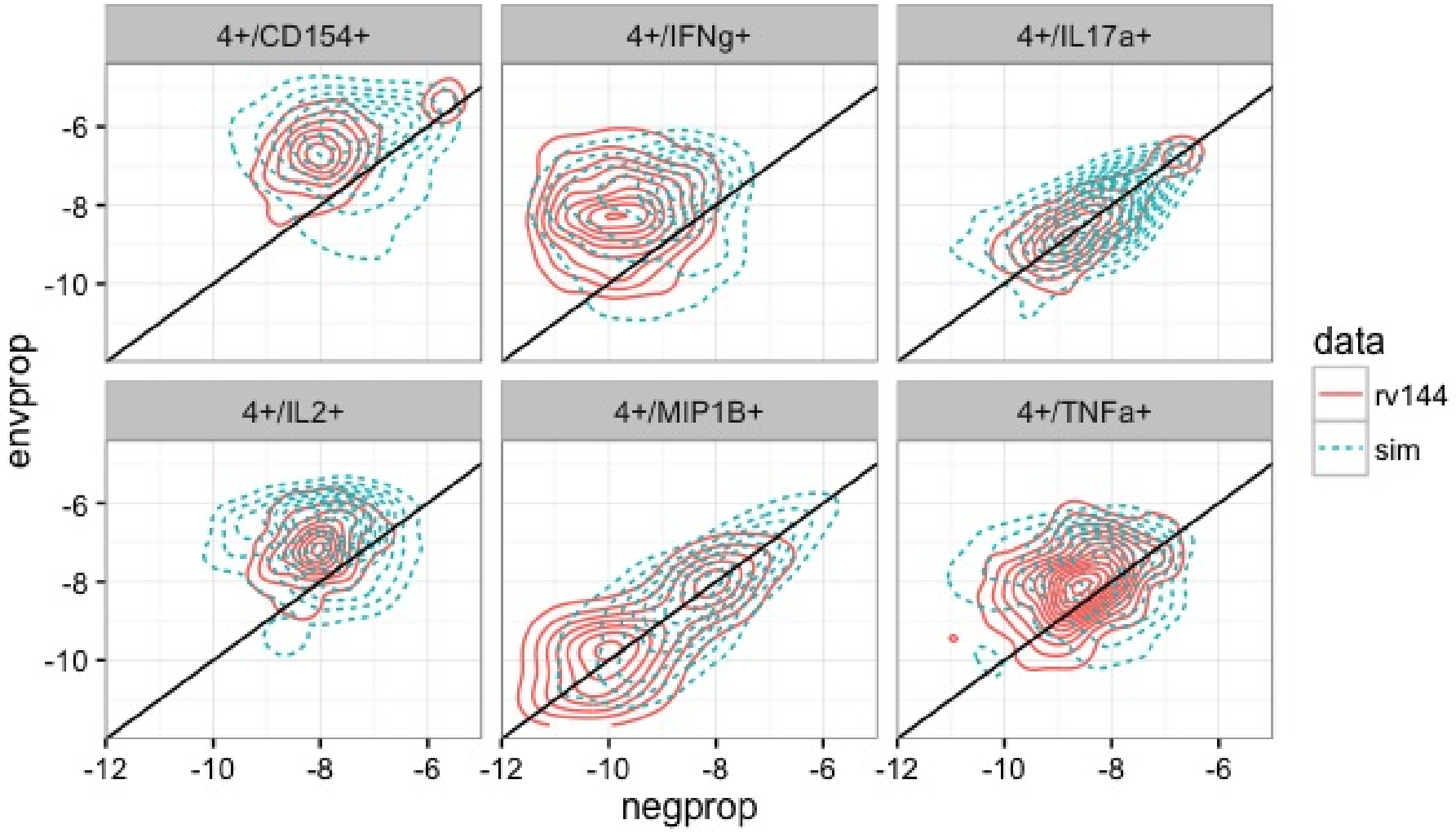
$$\nu_i \sim N(0, \Sigma),$$

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

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

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

How close are we to the distribution of the data?



Overdispersed Model - Results

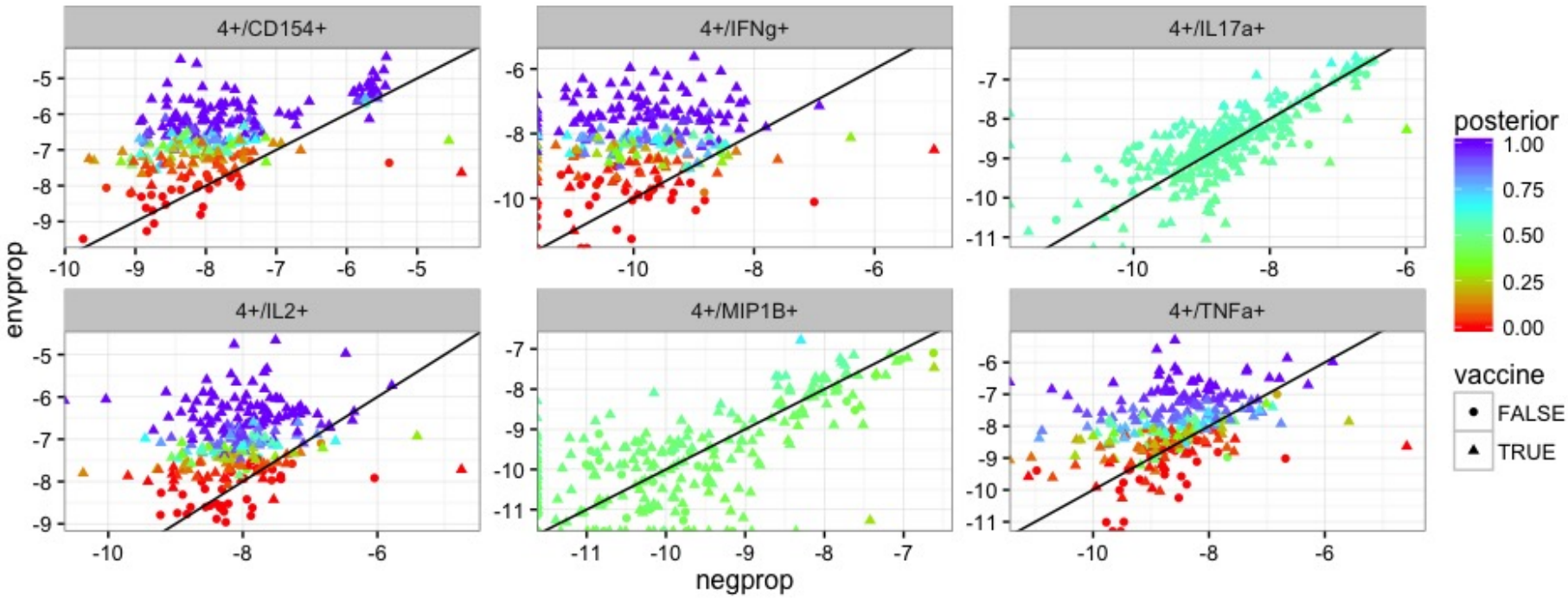


Figure: Scatter plot for Overdispersed Model

Overdispersed Model - Results

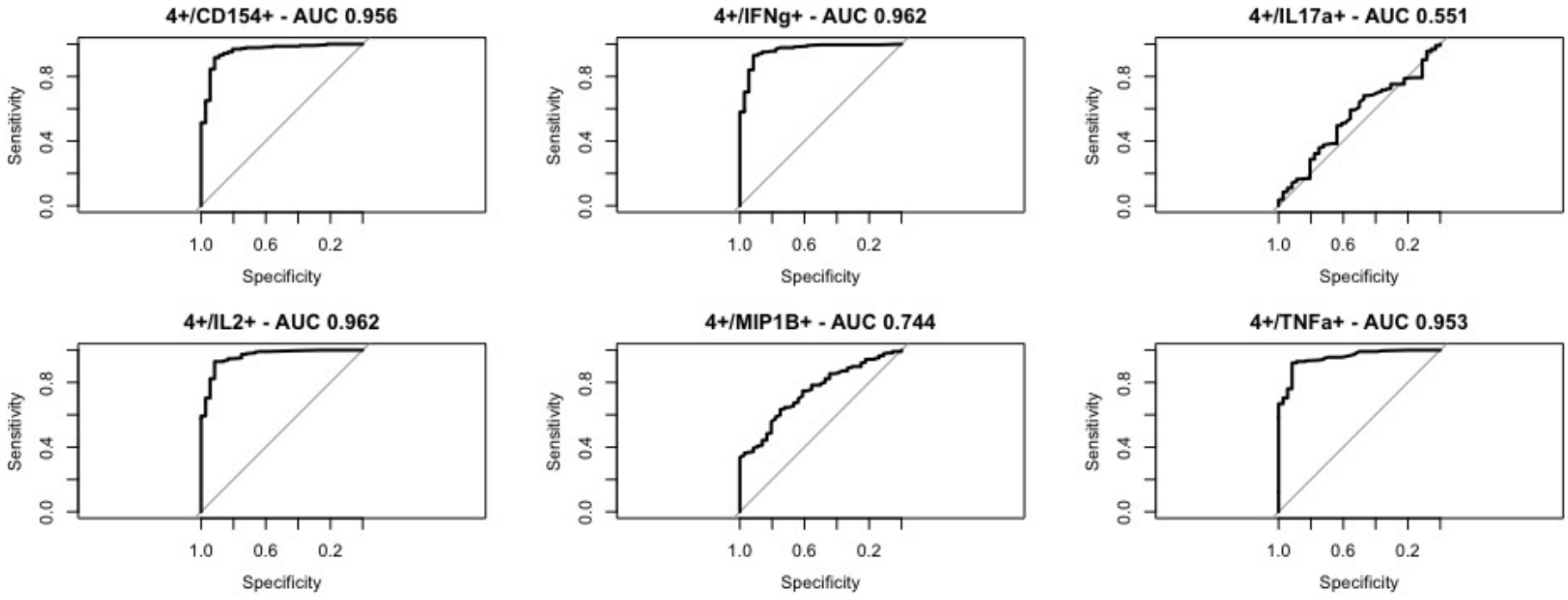


Figure: ROC for Overdispersed Model

Overdispersed Model - Results

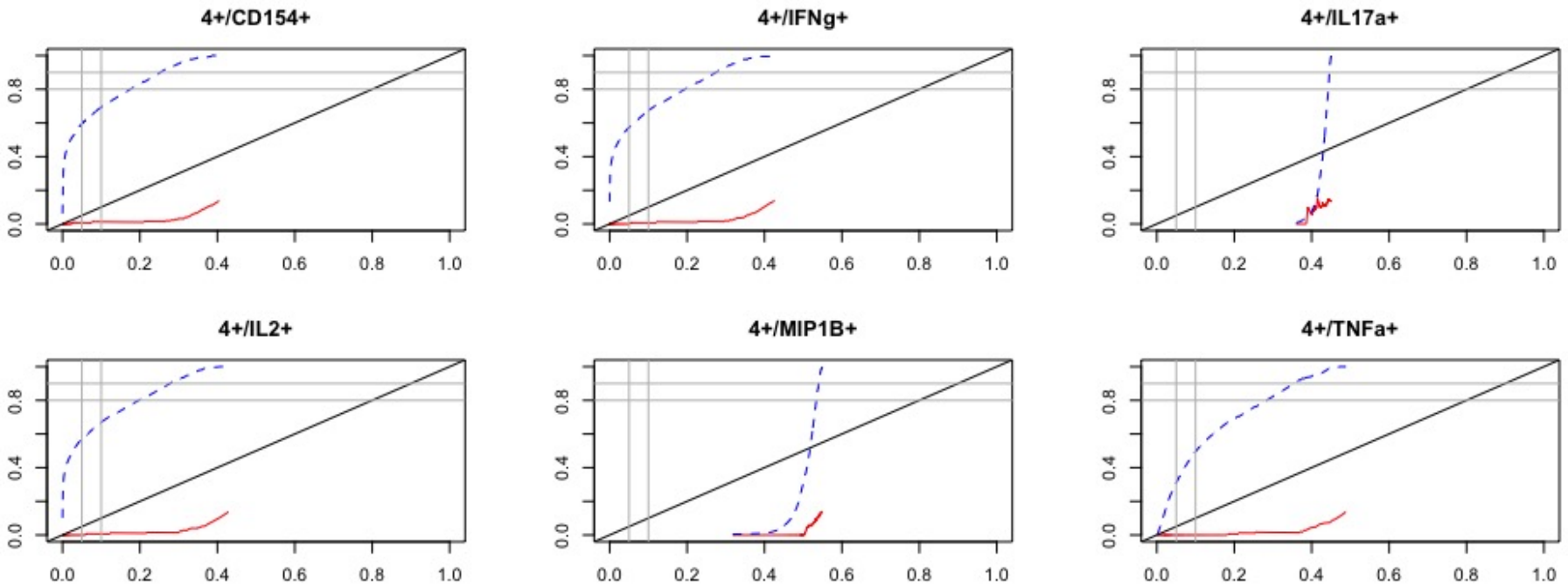
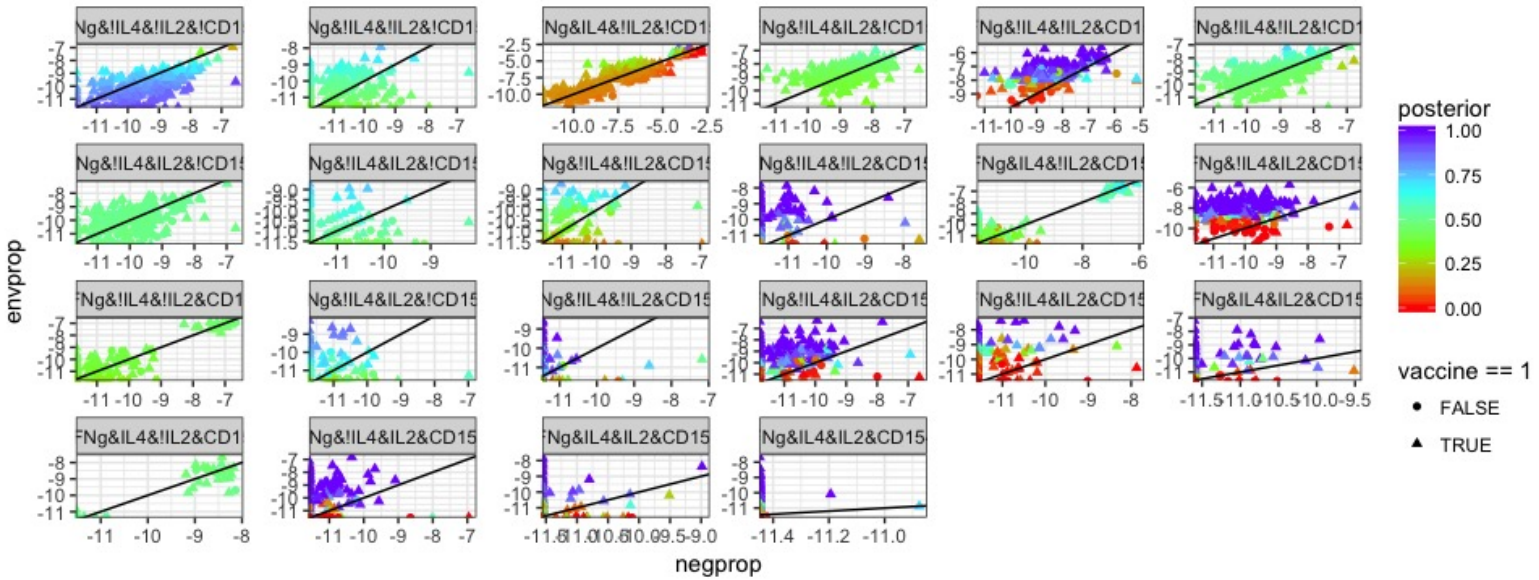


Figure: FDR for Overdispersed Model

RV144 - Booleans Dataset

226 vaccinees, and 36 placebos, 24 cell-subsets.



RV144 - Booleans Dataset

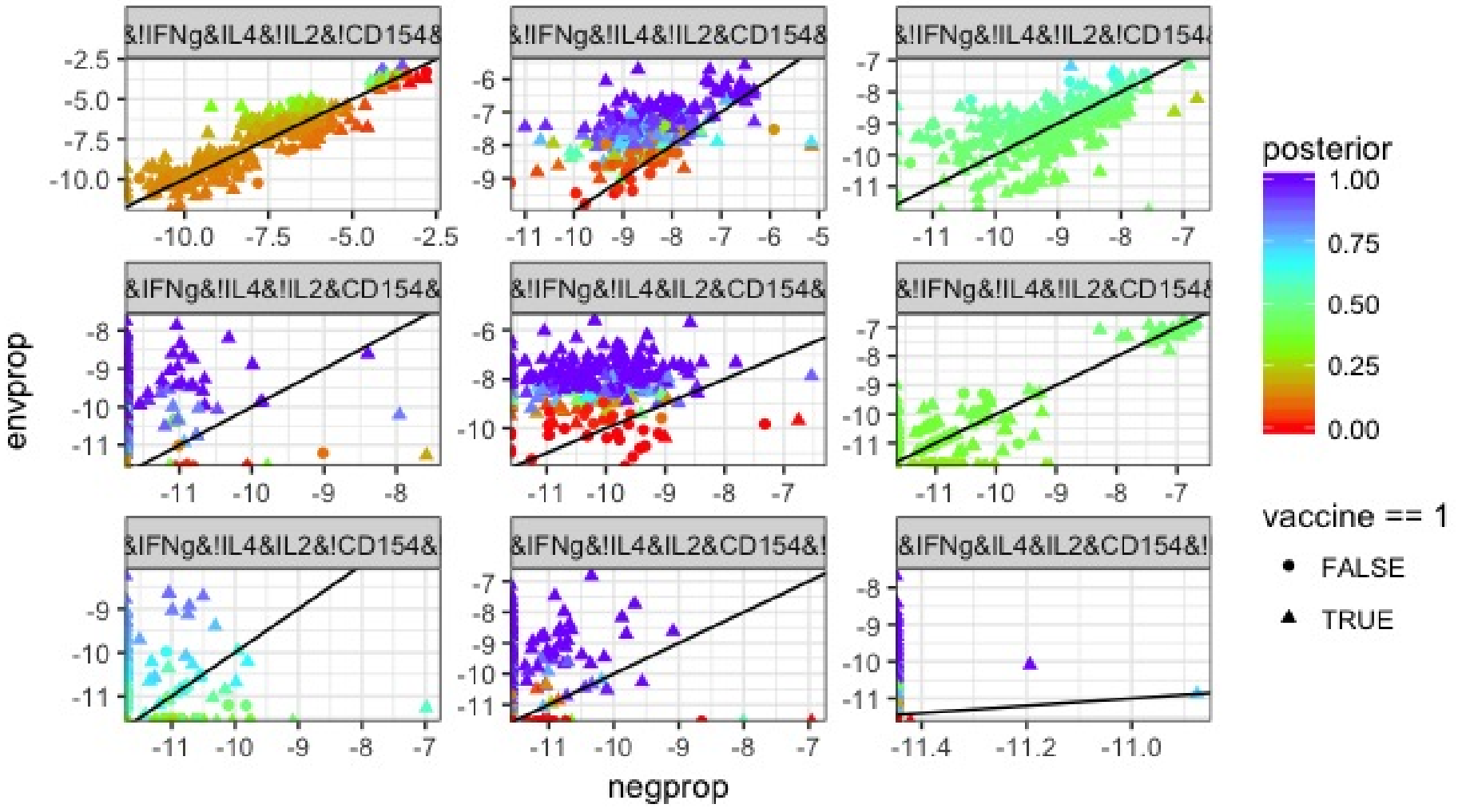


Figure: Scatter plots for RV144 booleans dataset

RV144 - Booleans Dataset

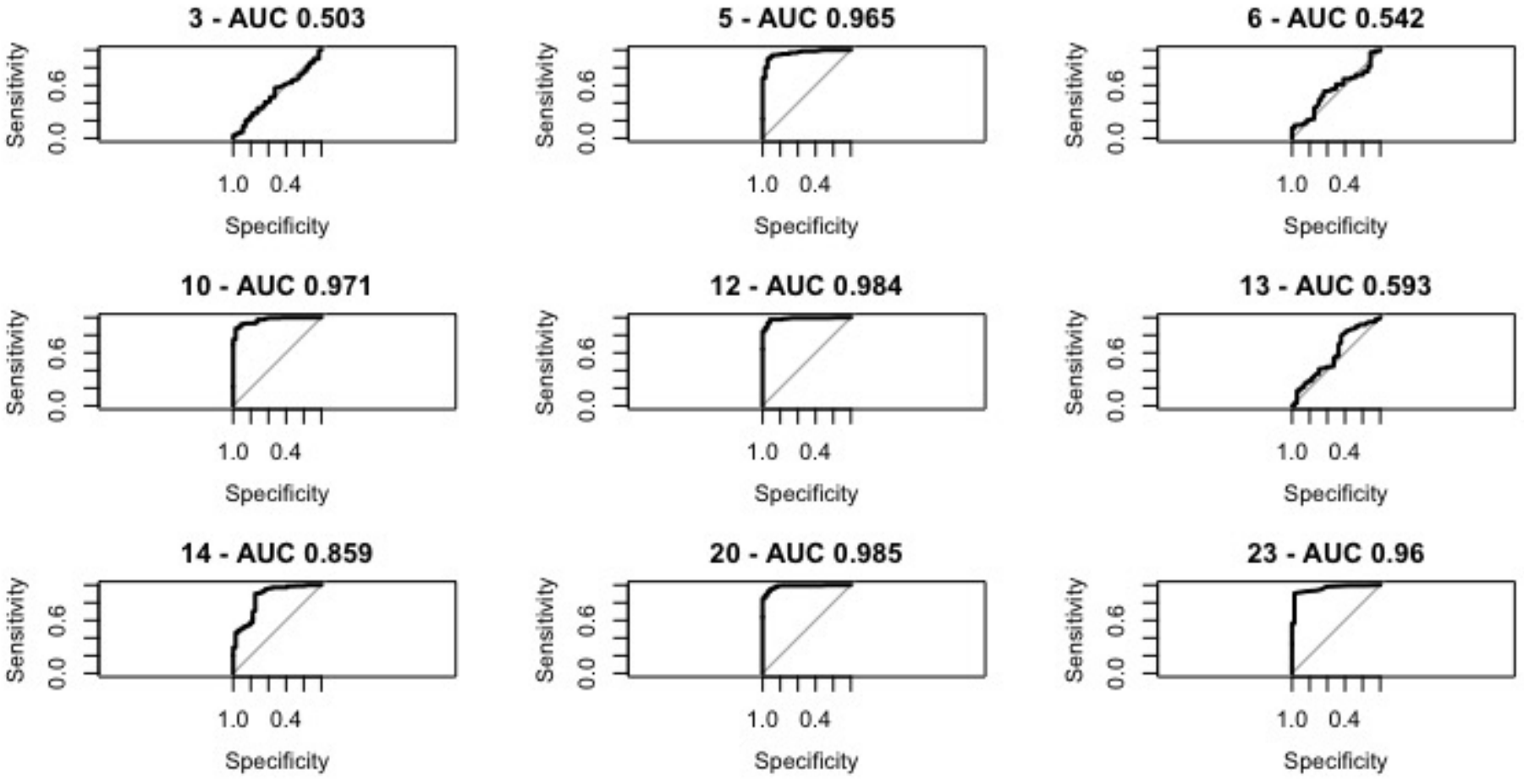


Figure: ROC for RV144 booleans dataset

RV144 - Booleans Dataset

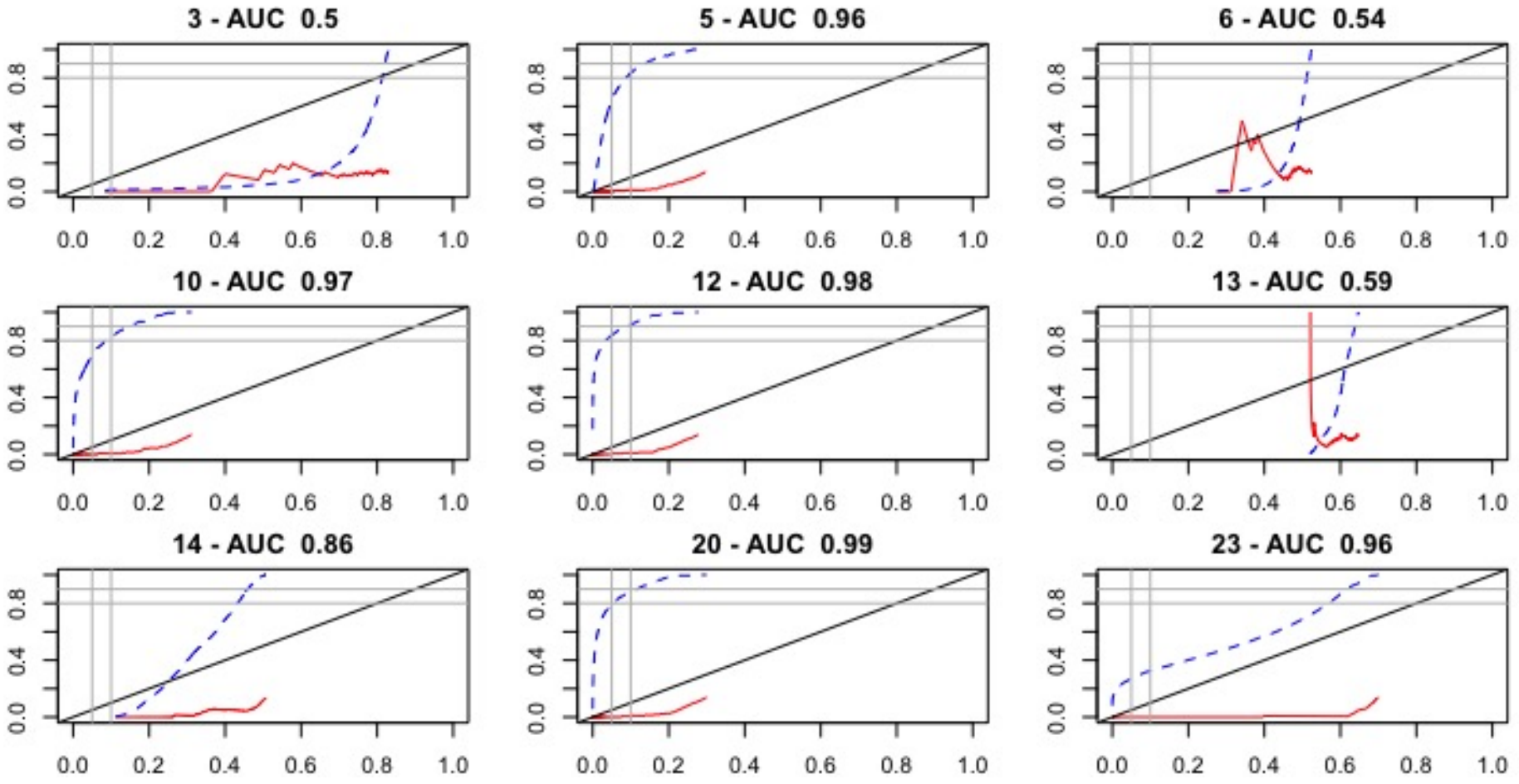


Figure: FDR for RV144 booleans dataset

RV144 - Booleans Dataset

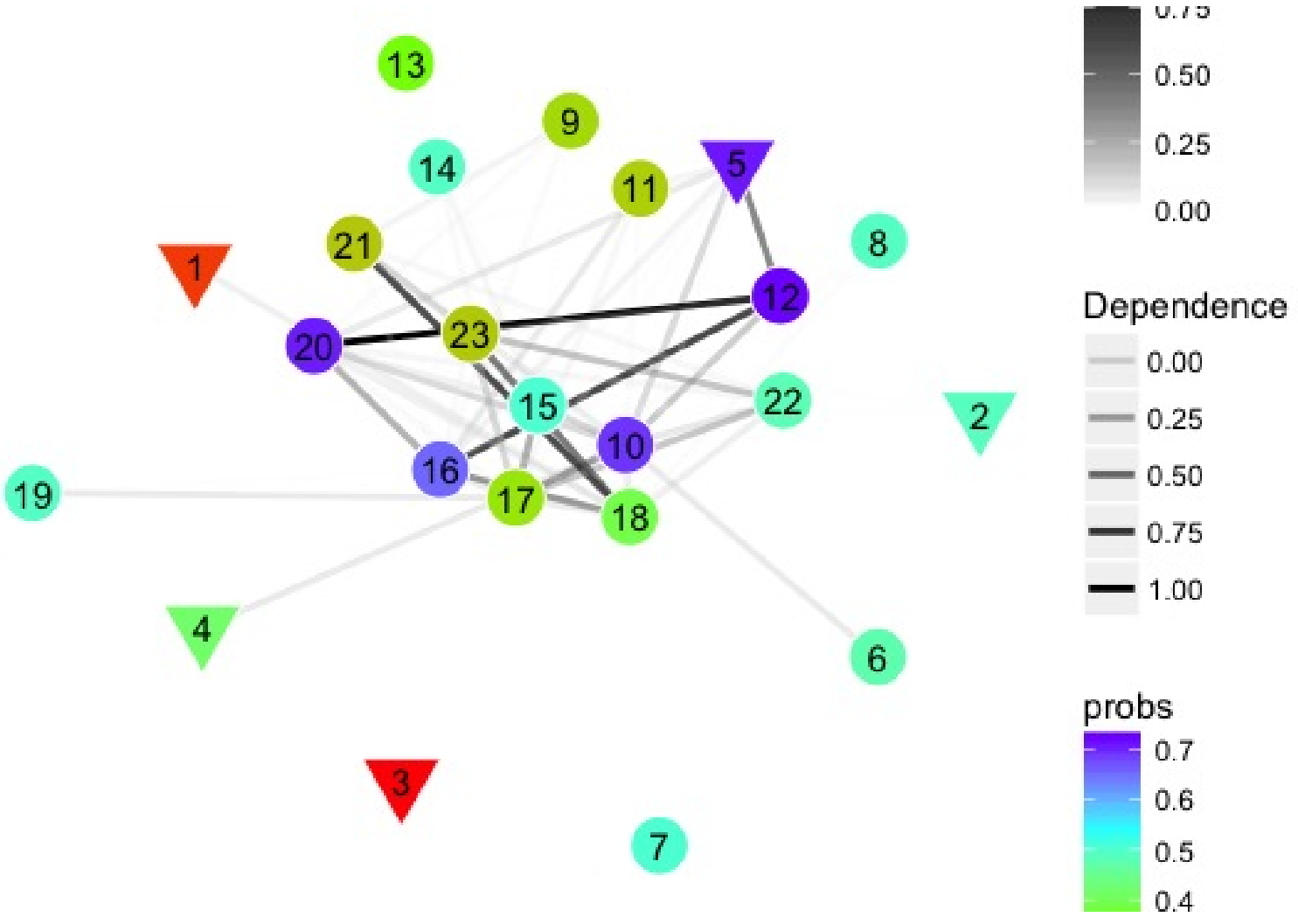


Figure: Estimated Ising Model

RV144 - Booleans Dataset

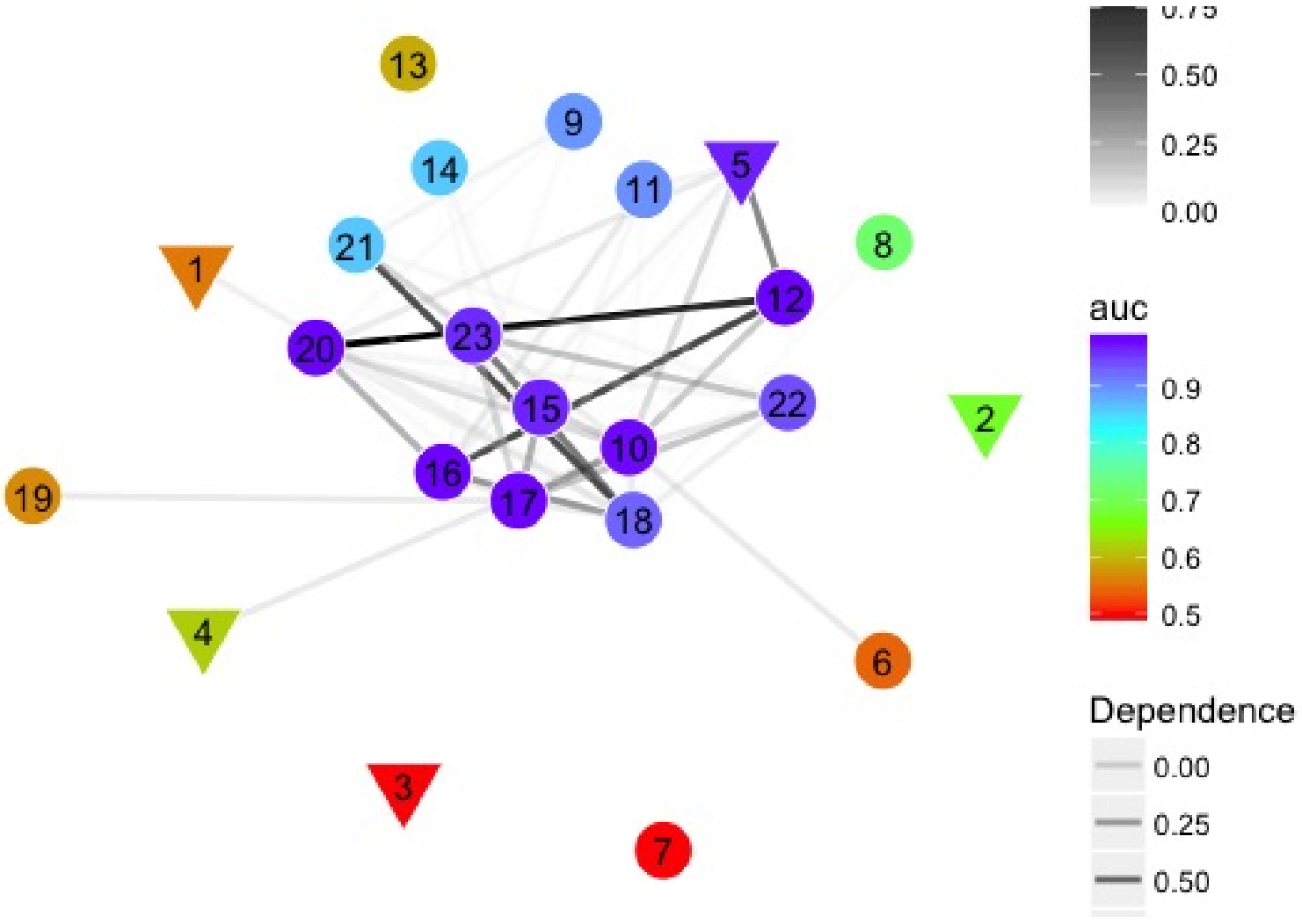


Figure: Estimated Ising Model

Whats next...

More data analysis!

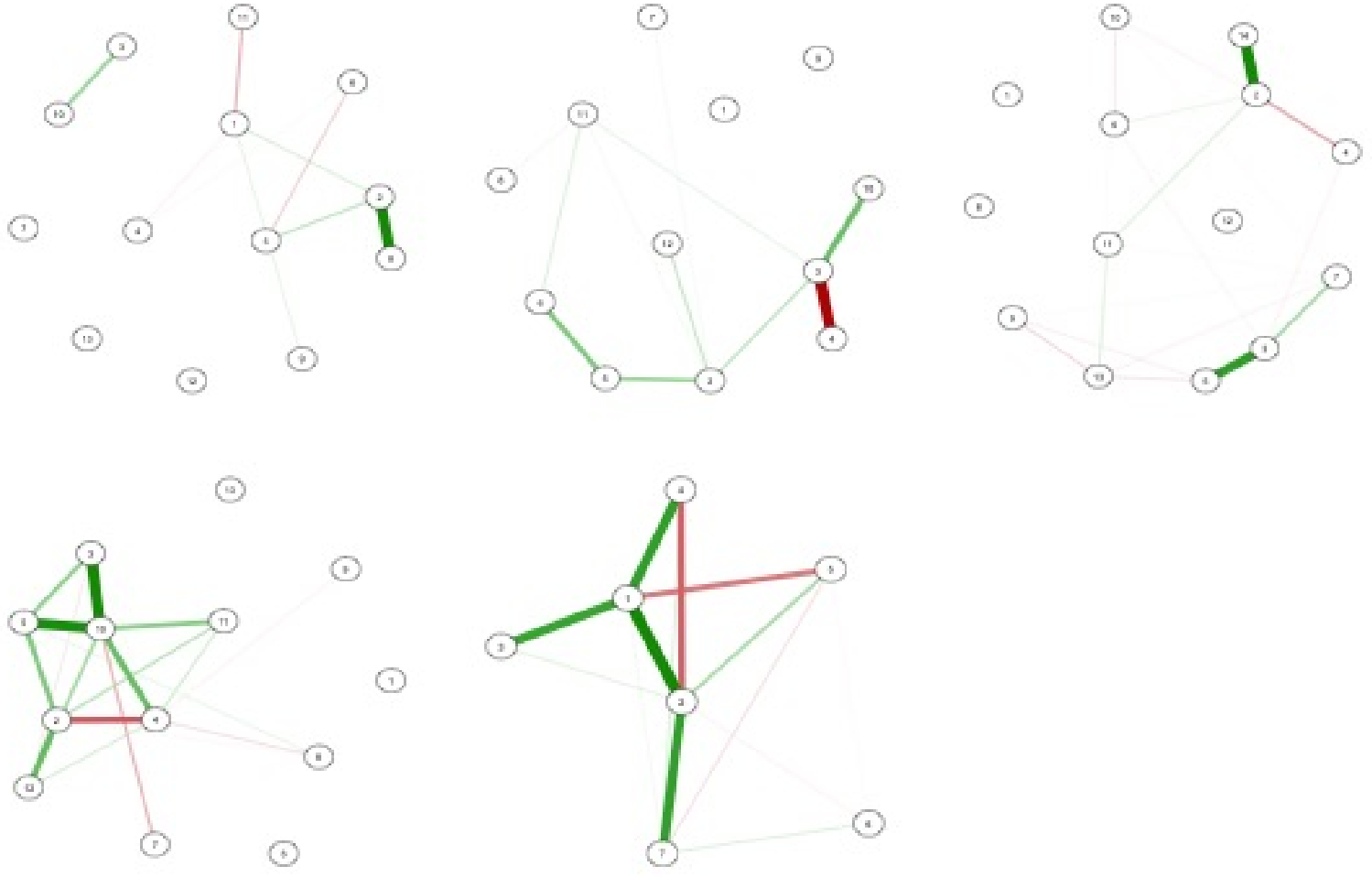


Figure: Estimated graphs from malaria dataset

Thank you!

Questions?

AmitMeir@uw.edu