flowReMix
(temporary name)

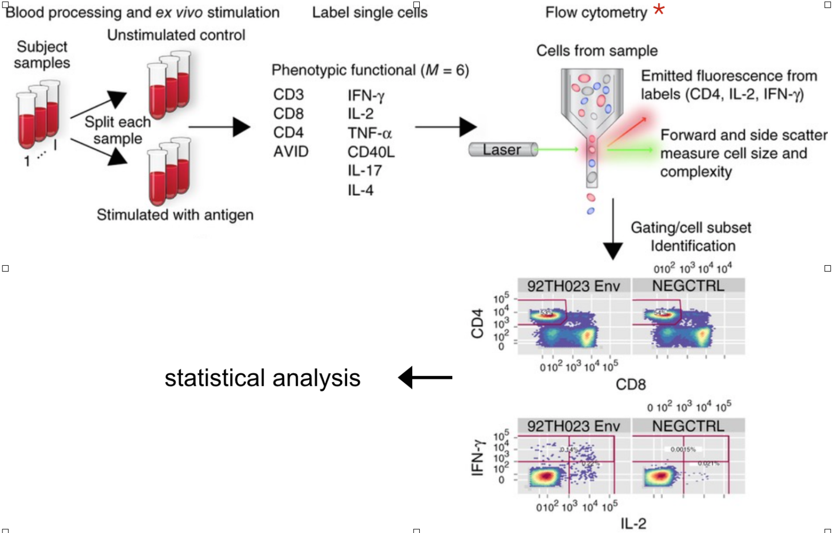
A **Mixture** of **Mixed** Beta-Binomial **Regression** Models for Analyzing **Flow**-Cytometry Count data

February 20, 2017

Outline

1. Introduction to Flow-Cytometry
2. Motivation
3. Models:
   - A Marginal Model
   - A Joint HMRF model
4. Data analysis.
5. Computation

Introduction to Cytometry Count Data



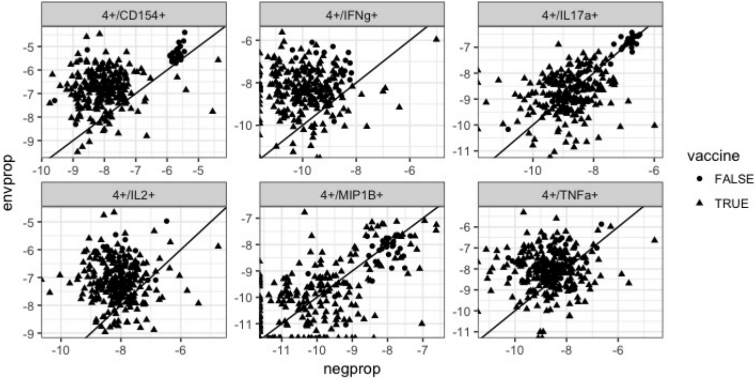
statistical analysis ←

adapted from Lin et al, Nat. Biotech. 2015

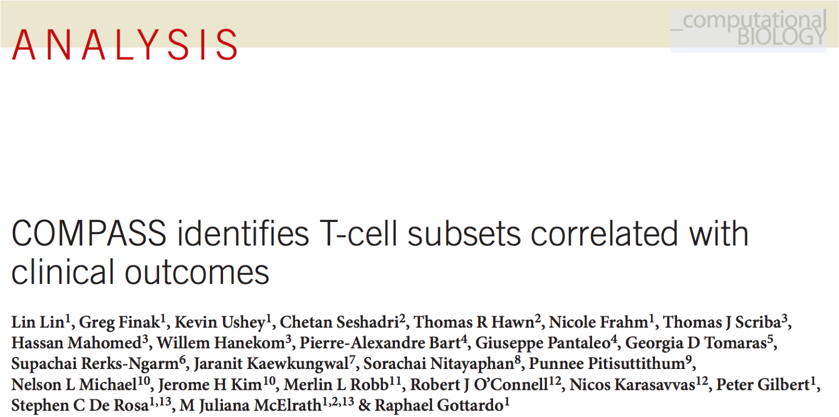
The RV144 HIV Vaccine Study

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

Marginal Counts for RV144



Motivation: COMPASS



ANALYSIS

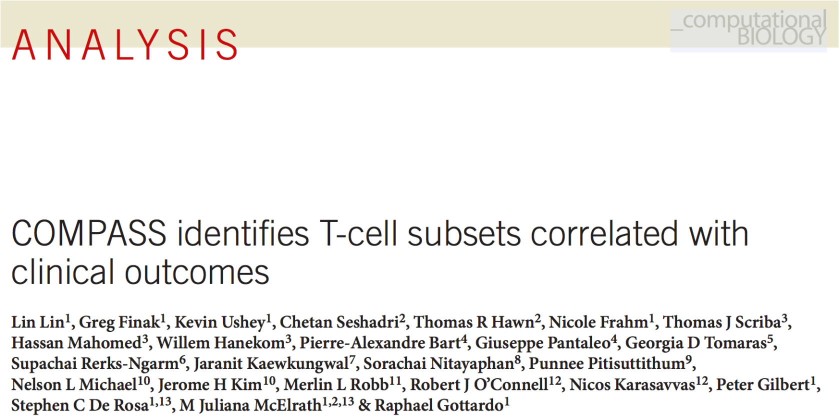
computational BIOLOGY

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.
- Goal is to identify subsets that respond to stimulation.

For the *i*th subject out of *n*,

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

Let $\tau_i \in \{0, 1\}^p$ be a vector indicating whether a subject shows response in one of the *p* cell subsets:

$$\tau_i \sim Bin(w),$$

$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0})$$

$$p_{i1,\tau=1}|p_{i0,\tau=0} \propto Dir(\alpha_1, \beta_1)$$

How do current methods work? (Approximately)

Current models are baseline/stimulation models.

- Unstimulated blood sample are compared stimulated ones.
- Goal is to identify subsets that respond to stimulation.

For the *i*th subject out of *n*,

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

Let $\tau_i \in \{0, 1\}^p$ be a vector indicating whether a subject shows response in one of the *p* cell subsets:

$$\tau_i \sim Bin(w),$$

$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0})$$

$$p_{i1,\tau=1} | p_{i0,\tau=0} \propto Dir(\alpha_1, \beta_1)$$

How do current methods work? (Approximately)

Current models are baseline/stimulation models.

- Unstimulated blood sample are compared stimulated ones.
- Goal is to identify subsets that respond to stimulation.

For the *i*th subject out of *n*,

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

Let $\tau_i \in \{0, 1\}^p$ be a vector indicating whether a subject shows response in one of the *p* cell subsets:

$$\tau_i \sim Bin(w),$$

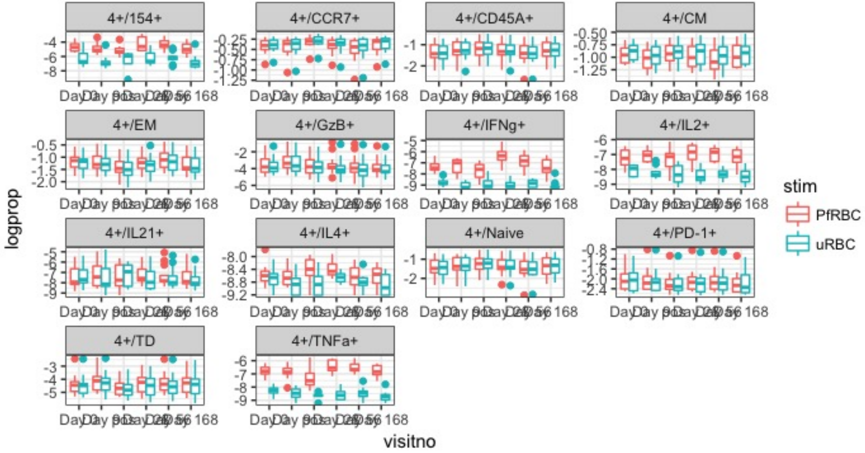
$$p_{i1,\tau=0} \sim \delta(p_{i0,\tau=0})$$

$$p_{i1,\tau=1} | p_{i0,\tau=0} \propto Dir(\alpha_1, \beta_1)$$

Controlled Human Malaria Infection Study

- 9 Tanzanian adults 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.
- 113 measured cell-types divided into 8 groups.

Controlled Human Malaria Infection Study



Motivation - Unique Challanges

- **Dependence**
  - Within sample between cell subsets.
  - Within subject / across time.

- **Heterogenous treatment effect**

- **Over-dispersed Binomial counts**

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.

A Marginal Model - Single Subset

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

- Binomial count data ⇒ Logistic model.

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

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

- Dependence ⇒ ‘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/time-point.

- Binomial count data ⇒ Logistic model.

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

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

- Dependence ⇒ '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 $k = 1$.

Model can be estimated via an EM algorithm

A Marginal Model - Single Subset

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

- Non-response ⇒ 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 $k = 1$.

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

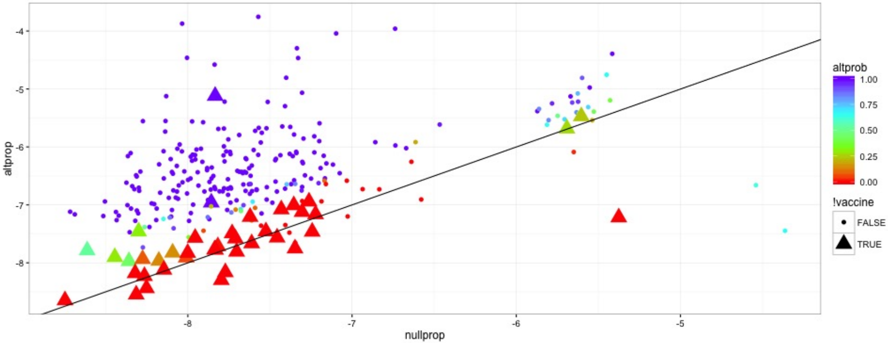


Figure: Scatter plot for T4+/CD154+ - Marginal Model

Marginal Model - Results

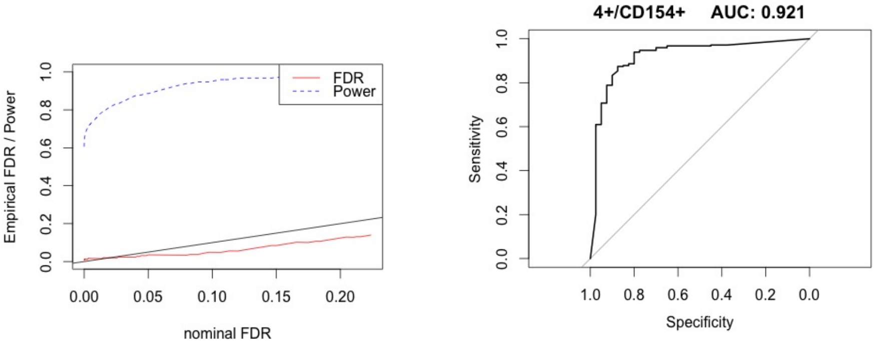


Figure: ROC/FDR plots for T4+/CD154+ - Marginal Model

Finak et al. (2013) - MIMOSA

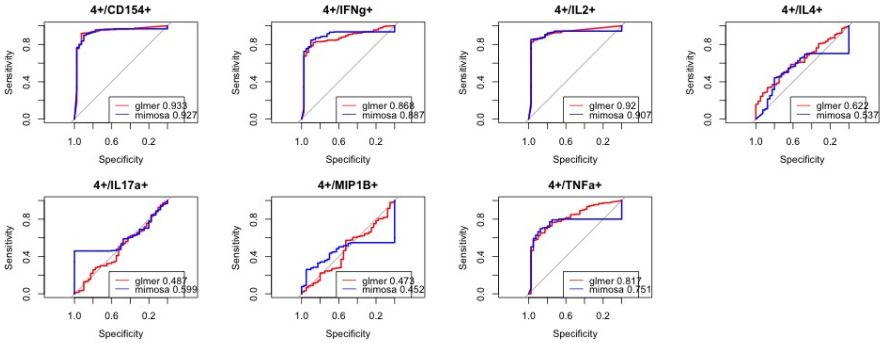


Figure: Comparison with MIMOSA (univariate COMPASS)

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 very likely 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 very likely 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 very likely 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, **l**- 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, **l**- 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, **l**- 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, **l**- stimulation, **j**- subset, **k**- cluster.

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

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

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

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

HMRF Modle - Results

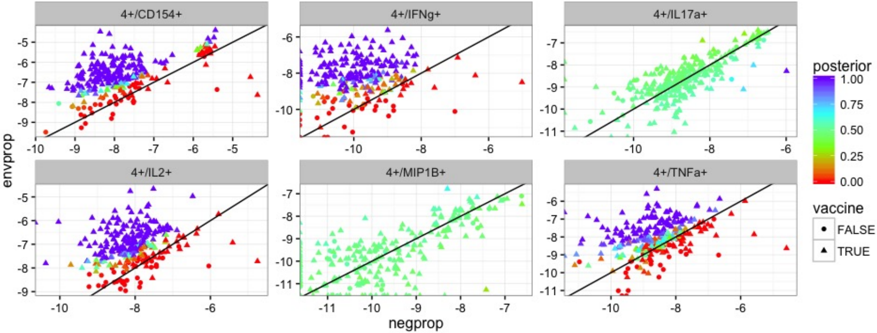


Figure: Scatter Plot for HMRF Modle Model

Subset-Response Model - Results

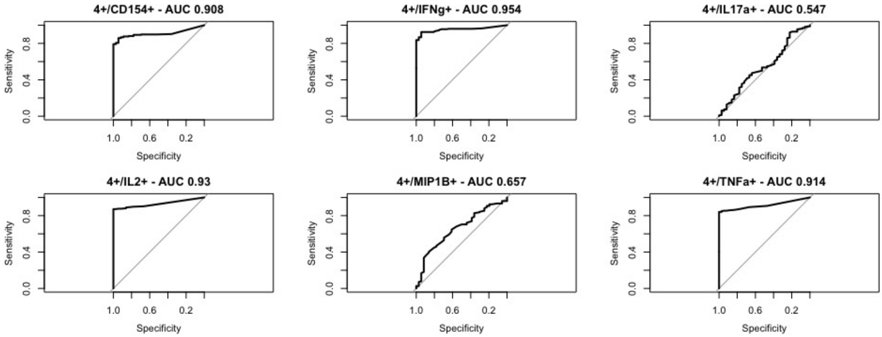


Figure: ROC for HMRF Modle

Subset-Response Model - Results

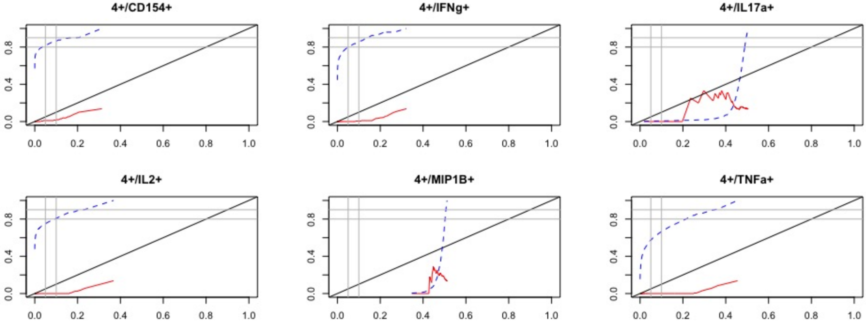


Figure: FDR for HMRF Modle

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?

Simulated Binomial Data - Results

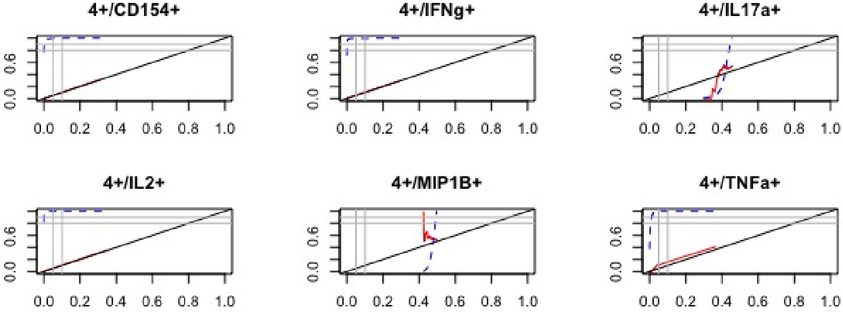


Figure: FDR for Simulated Binomial Data

Simulated Binomial Data - Results

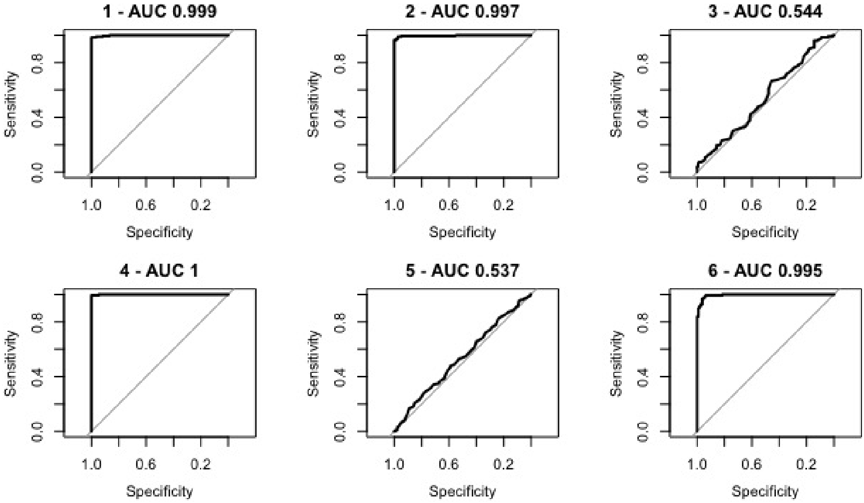


Figure: ROC for Simulated Binomial Data

Simulated Binomial Data - Results

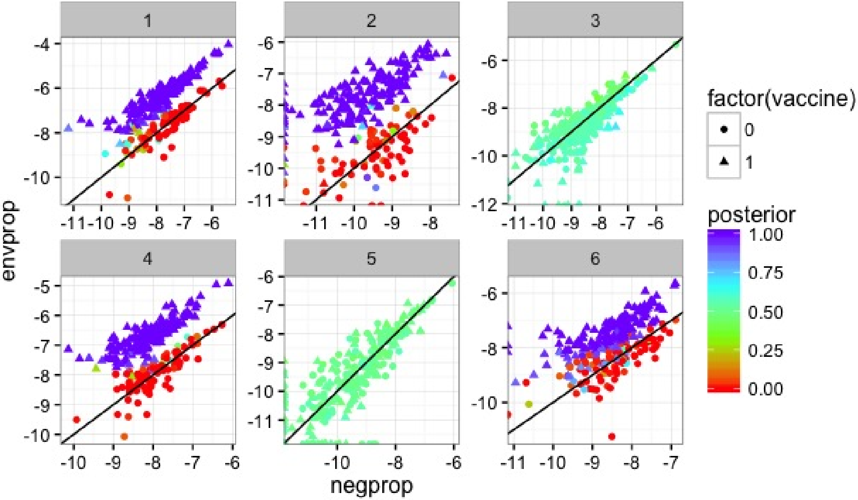


Figure: Scatter plot for Simulated Binomial Data

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.

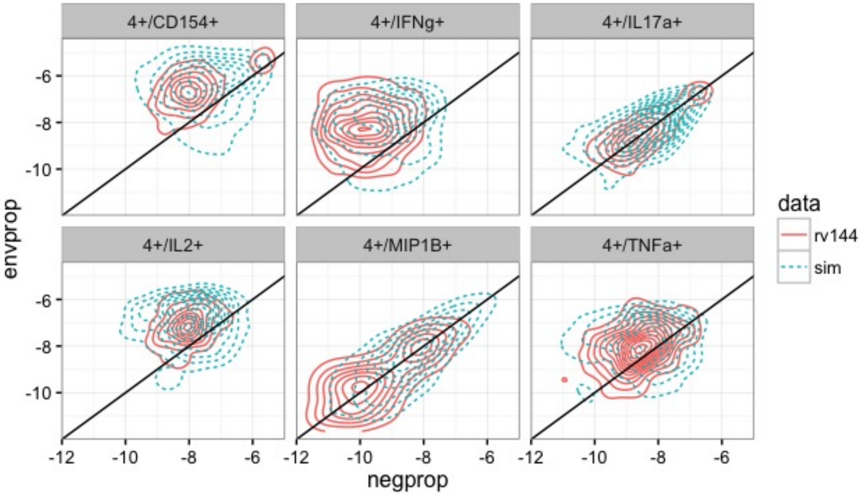
$$\nu_i \sim N(0, \Sigma),$$

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

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

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

How close are we to the distribution of the data?



Overdispersed Model - Results

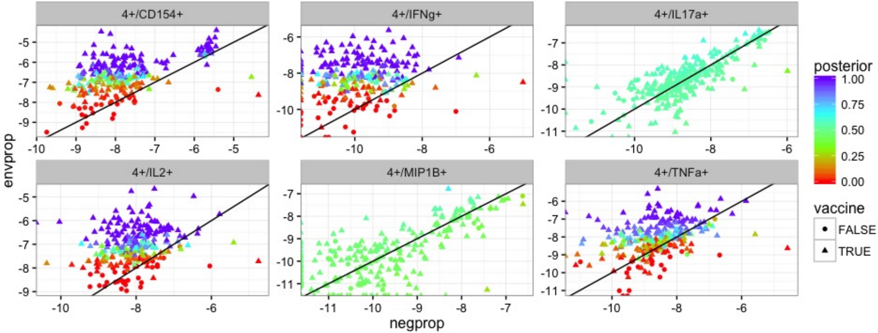


Figure: Scatter plot for Overdispersed Model

Overdispersed Model - Results

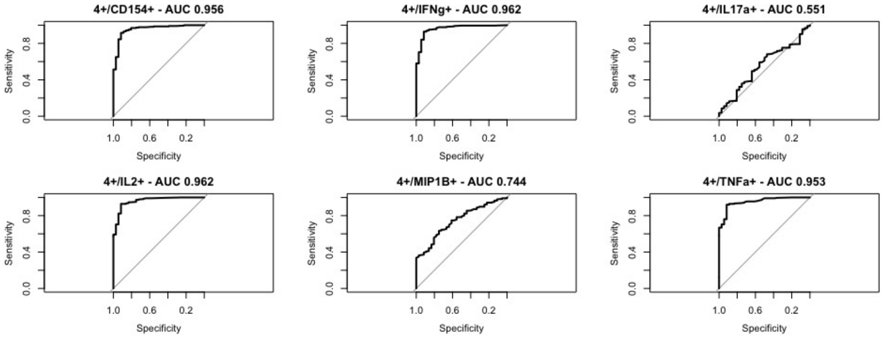


Figure: ROC for Overdispersed Model

Overdispersed Model - Results

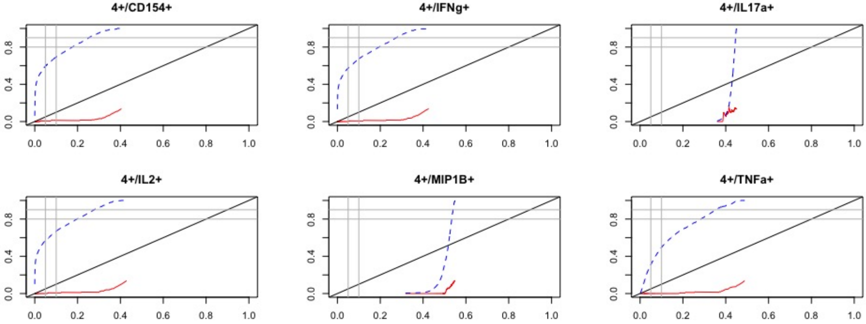
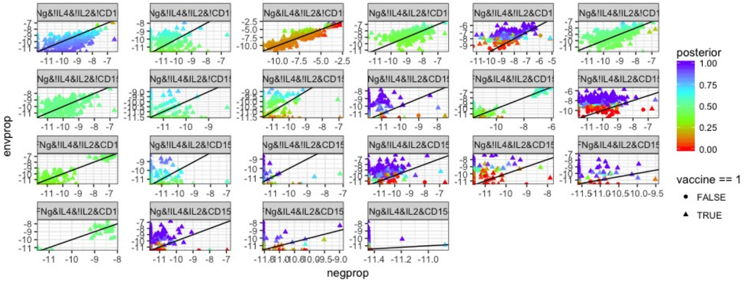


Figure: FDR for Overdispersed Model

RV144 - Booleans Dataset

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



RV144 - Booleans Dataset

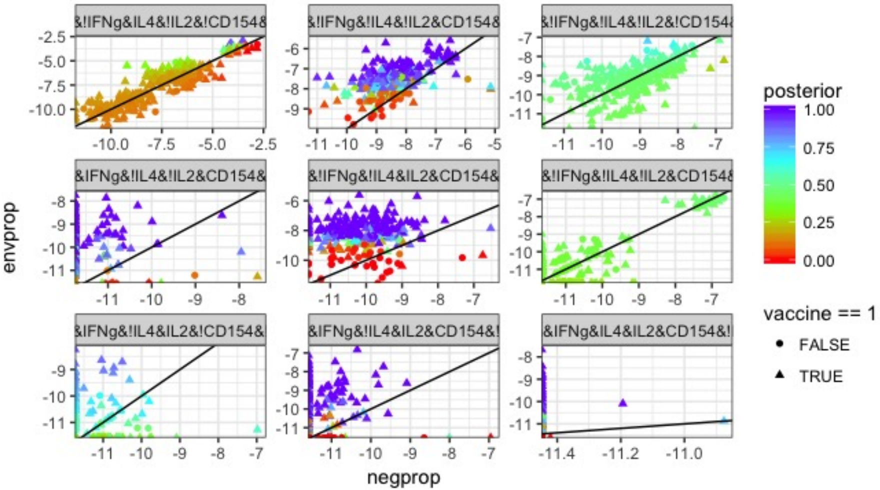


Figure: Scatter plots for RV144 booleans dataset

RV144 - Booleans Dataset

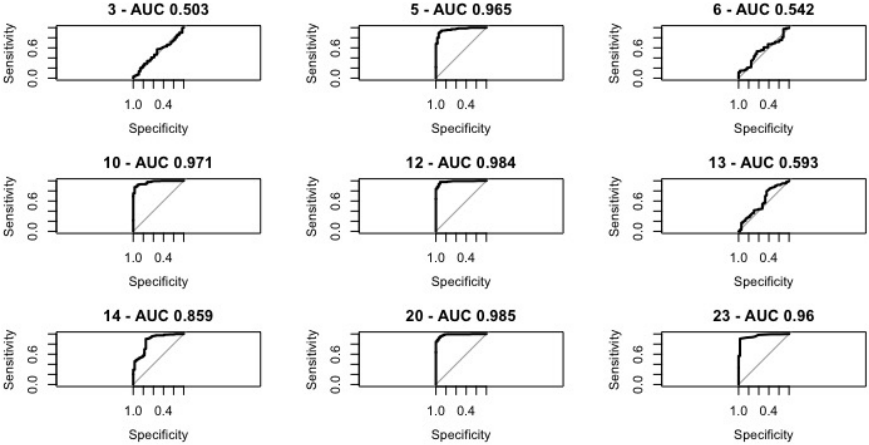


Figure: ROC for RV144 booleans dataset

RV144 - Booleans Dataset

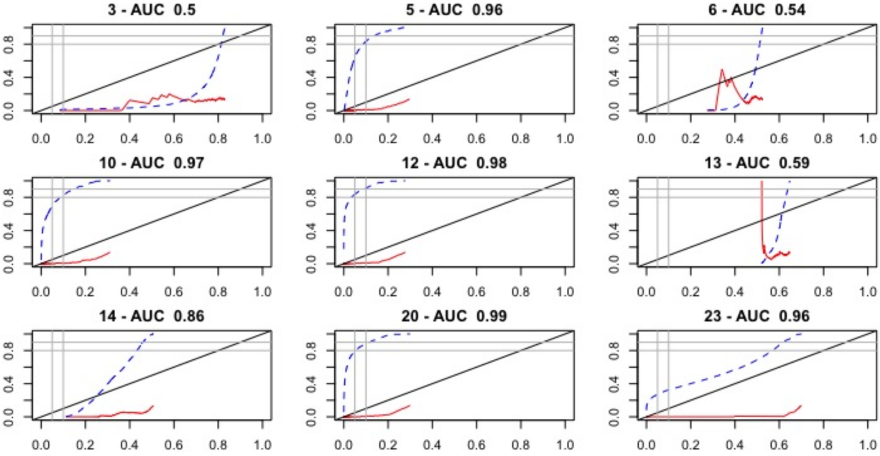


Figure: FDR for RV144 booleans dataset

RV144 - Booleans Dataset

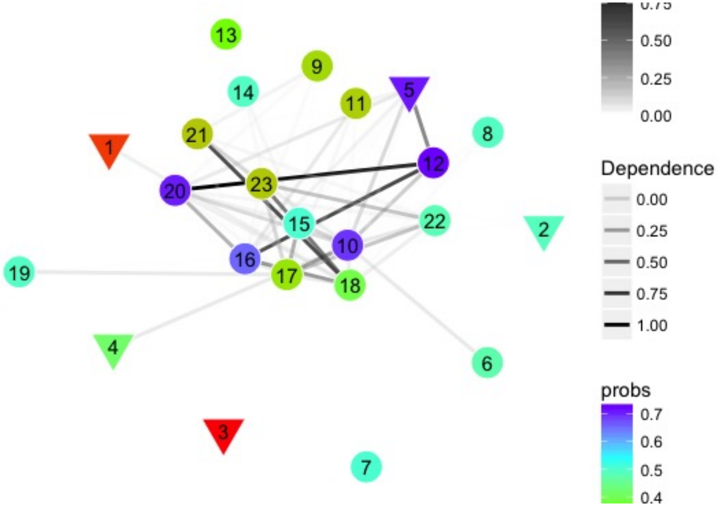


Figure: Estimated Ising Model

RV144 - Booleans Dataset

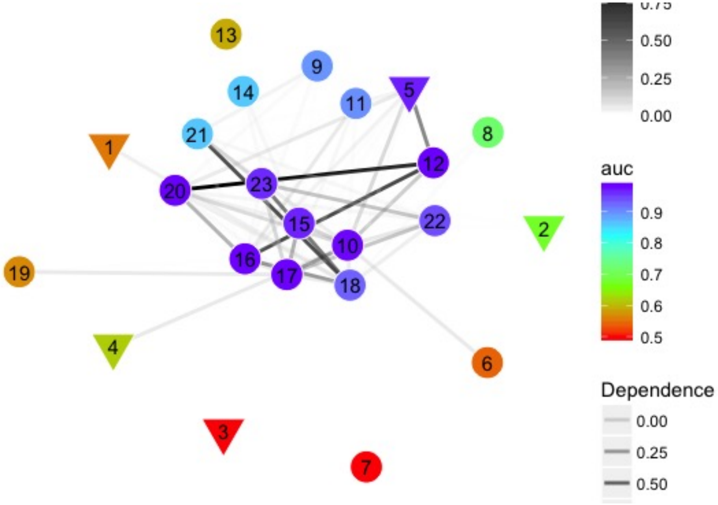


Figure: Estimated Ising Model

Controlled Human Malaria Infection Study

- 9 Tanzanian adults 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.
- 113 measured cell-types divided into 8 groups.

Controlled Human Malaria Infection Study

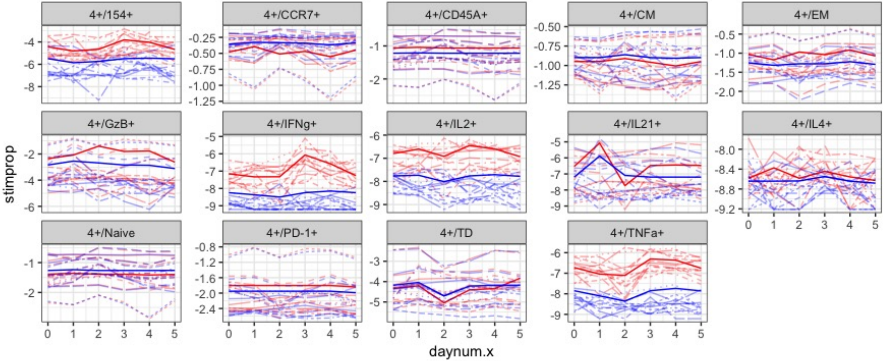
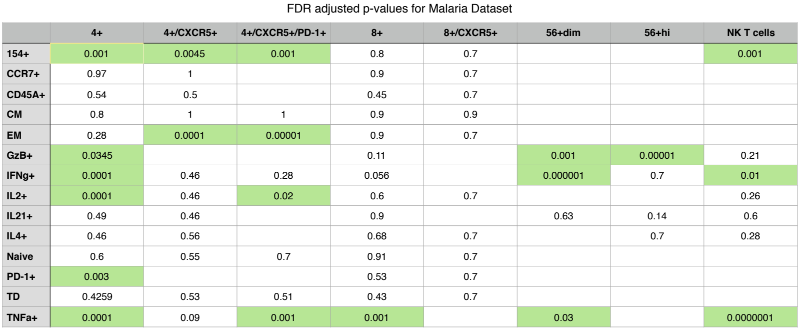


Figure: Stimulated vs. Unstimulated

Controlled Human Malaria Infection Study



FDR adjusted p-values for Malaria Dataset

	4+	4+/CXCR5+	4+/CXCR5+/PD-1+	8+	8+/CXCR5+	56+dim	56+hi	NK T cells
154+	0.001	0.0045	0.001	0.8	0.7			0.001
CCR7+	0.97	1		0.9	0.7			
CD45A+	0.54	0.5		0.45	0.7			
CM	0.8	1	1	0.9	0.9			
EM	0.28	0.0001	0.00001	0.9	0.7			
GzB+	0.0345			0.11		0.001	0.00001	0.21
IFNg+	0.0001	0.46	0.28	0.056		0.000001	0.7	0.01
IL2+	0.0001	0.46	0.02	0.6	0.7			0.26
IL21+	0.49	0.46		0.9		0.63	0.14	0.6
IL4+	0.46	0.56		0.68	0.7		0.7	0.28
Naive	0.6	0.55	0.7	0.91	0.7			
PD-1+	0.003			0.53	0.7			
TD	0.4259	0.53	0.51	0.43	0.7			
TNFa+	0.0001	0.09	0.001	0.001		0.03		0.0000001

Thank you!

Questions?

AmitMeir@uw.edu