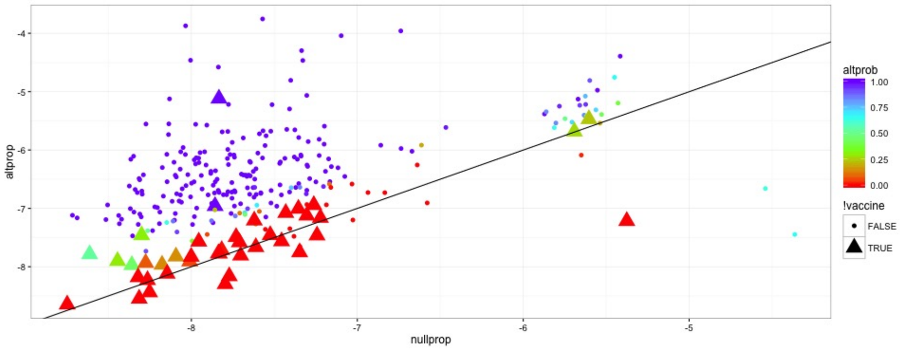

Figure 1: Scatter plot of the log-proportions for the T4+/CD154 cell-subset. The x-axis is the log proportion in the control sample and the y-axis is the log proportion in the stimulated sample. Subject in the control group are marked with triangles and subject in treatment group as circles. The posterior probability of being a responder (or in the treatment group) is marked by color, purple being a probability close to one and red a posterior probability close to zero.

1 A Mixture of Mixed GLMs

1.1 Mixed Binomial Regression

Cell count data exhibits over-dispersion when compared to what would be expected from a standard binomial experiment. This can be seen in Figure 1 where the diagonal line serves as a baseline proportion for each subject. Subjects in the control group tend to be close to the diagonal line and subjects in the treatment group tend to be above the diagonal line (because the stimulated sample has a higher count of cells than the unstimulated sample).

While the Beta-Binomial distribution is a good model for over-dispersion when the over-dispersion is observed at the blood-sample level, here most of the over-dispersion seems to be at the subject-level and when this variability is accounted for the over-dispersion observed at the sample level is actually very small. We can model the count of a specific blood cell type in the jth sample from the ith subject with the following model:

$$\begin{aligned} y_{ij} &\sim \mathrm{Bin}(N_{ij}, p_{ij}), \\ \mathrm{logit}(p_{ij}) &= X_{ij}\beta + \nu_i, \\ \nu_i &\sim N(0, \sigma^2). \end{aligned}$$

Giving rise to the likelihood:

$$f(y_i) = \int_{\mathbb{R}} \prod_j f(y_{ij}; X_{ij}, \beta, \nu)\varphi(\nu; \sigma^2) d\nu.$$

There are many existing software packages for fitting such regression models which work well as long as the dimension of the integral is not too big (≤ 3).

1.2 A Finite Mixture Of Mixed Regression

In addition to accounting for over-dispersion we must also account for the fact that we may have non-responders in our study. To account for that we can use the following model:

$$y_{ij} \sim \pi_0 \mathrm{Bin}(N_{ij}, p_{ij0}) + \pi_1 \mathrm{Bin}(N_{ij}, p_{ij1}),$$
$$\pi_0 + \pi_1 = 1,$$
$$\mathrm{logit}(p_{ijk}) = X_{ij}\beta + T_{ij}\tau_k + \nu_i,$$
$$\nu_i \sim N(0, \sigma^2),$$
$$\tau_k = \begin{cases} \tau & k = 1 \\ 0 & k = 0 \end{cases}.$$

Where X_{ij} and T_{ij} are known matrices, τ_k is a random regression coefficient that equals either zero or an unknown constant (which needs to be estimated).

We can estimate this model using an SEM algorithm (Stochastic-EM). In the E-step we approximate the posterior probability of belonging to the kth cluster based on the observed data and current set of parameters:

$$\pi_{i0} = \frac{f_0(y_i)\pi_0}{f_0(y_i)\pi_0 + f_1(y_i)\pi_1},$$
$$\pi_{i1} = 1 - \pi_{i0},$$
$$f_k(y_i) = \int_{\mathbb{R}} \prod_j f(y_{ij}; X_{ij}, T_{ij}, \tau_k, \beta, \nu)\varphi(\nu; \sigma^2) d\nu.$$

In the M-step, we maximize the complete data log-likelihood:

$$\sum_{k=0}^{1} \sum_{i=1}^{n} \pi_{ik} \log f_k(y_i) + \pi_{ik} \log \pi_k,$$

given the posterior probabilities p_{ik} this can be done using a standard mixed modeling library (such as **lme4**).

The reason we need to use a Stochastic-EM algorithm rather than a standard EM is that in the in the E-step the posterior probabilities are not given in closed form and they must be approximated. We approximate them using a importance sampling scheme. For the ith subject and for $k \in \{0, 1\}$ we sample $v_{k1}, ...v_{kM} \sim N(\mu_{ik}, \sigma^2)$, where μ_{ik} is estimated as a part of the M-step. We then compute an approximation to the posterior probability:

$$\frac{f_0(y_i)\pi_0}{f_0(y_i)\pi_0 + f_1(y_i)\pi_1} = \left(1 + \frac{\pi_1}{\pi_0}\frac{f_1(y_i)}{f_0(y_i)}\right)^{-1}$$
$$\approx \left(1 + \frac{\pi_1}{\pi_0}\frac{\sum_{m=1}^{M} \frac{\varphi(v_{1m};0,\sigma^2)}{\varphi(v_{1m};\mu_{i1},\sigma^2)} \prod_j f_1(y_{ij}; v_m)}{\sum_{m=1}^{M} \frac{\varphi(v_{0m};0,\sigma^2)}{\varphi(v_{0m};\mu_{i0},\sigma^2)} \prod_j f_0(y_{ij}; v_m)}\right)^{-1} := w_{i0}.$$

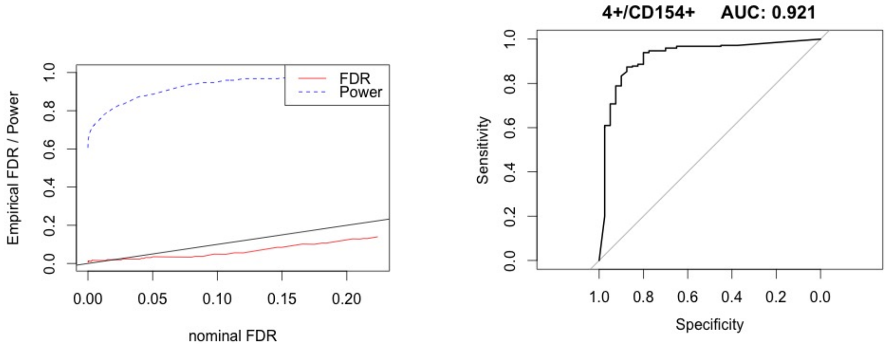

Figure 2: ROC and False Detection Rate for the T4+/CD154 cell subset.

Since w_{ik} is only a rough approximation for p_{ik}, we update the posterior probabilities gradually:

$$\pi_{ik}^{t} = \pi_{ik}^{t-1} + \frac{(w_{ik} - \pi_{ik}^{t-1})}{\sqrt{t}}.$$

2 Some Data Analysis

2.1 Modeling with a Single Random Effect

Let us fit this model to the RV144 dataset. Figure 1 describe the distribution of the log-proportions of T4+/CD154 cells in the two stimulus types and the posterior probabilities of being responders assigned by the algorithm to the subjects. The model assigns high probabilities to observations above the $x = y$ line and low probabilities to observations close to the line or below it. In Figure 2 we plot the ROC curve and nominal-FDR vs. empirical-FDR for the T4+/CD154 cell subset. The AUC is 0.921 and the False-Detection rate is controlled while maintaining good power.

In Figure 3 we plot the scatterplot, ROC and FDR for the T4+/IFNG+ cell-subset. Here the FDR is still good but the AUC is a lower of 0.87. This can be improved if we fit a model to the IFNG+ and CD154 cell-subsets together. At the moment, it is only possible to fit models with a single subject-level random effect. However, this works well enough in this case because, as we will show, the random effects in these two cell-subsets are highly correlated. The classification error and posterior probabilities given by the joint model are presented in Figure 4.

Joint modeling for multiple cell-subsets may fail if some of the random effects have different distributions. As we show in Figure 7, this is the case if we try to model the T4+/CD154 and T4+/IL4+ cell-subsets with one random effect. Here, the classification is worse than the one classification we would get had we fit a model for the T4+/CD154 cell subset alone.

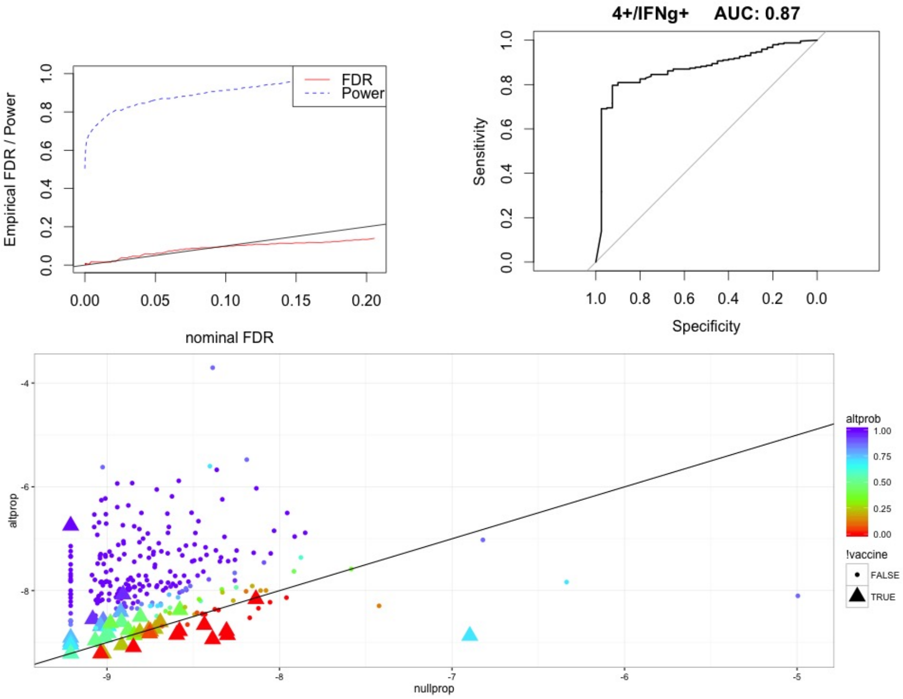

Figure 3: Plots for the T4+/IFNG+ cell subset.

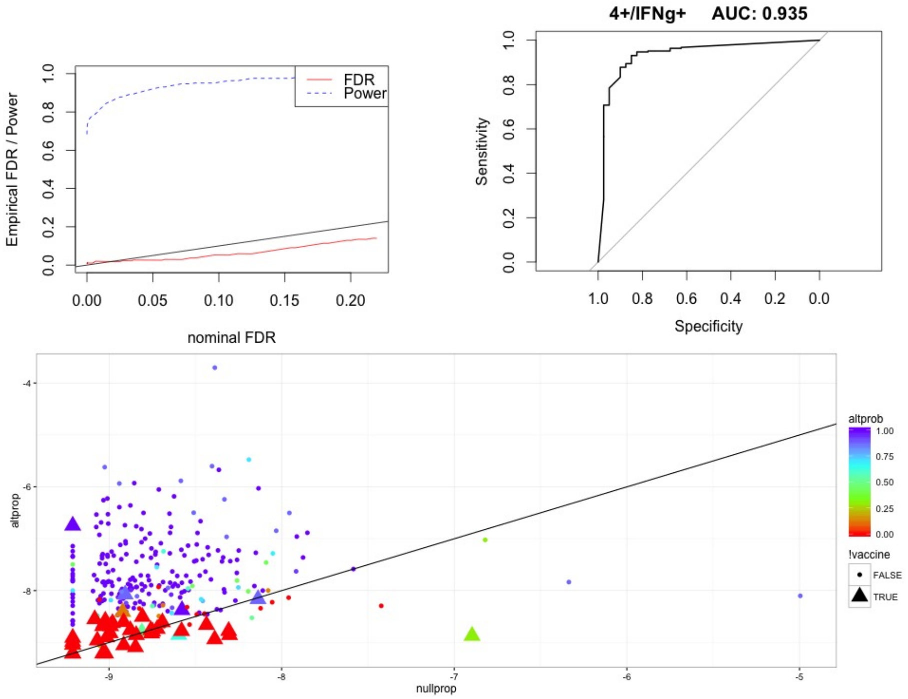

Figure 4: Posterior probabilities and classification errors for the joint model of T4+/CD154 and T4+/IFNG+. The scatter plot is for log proportions of the T4+/IFNG+ cell-subset.

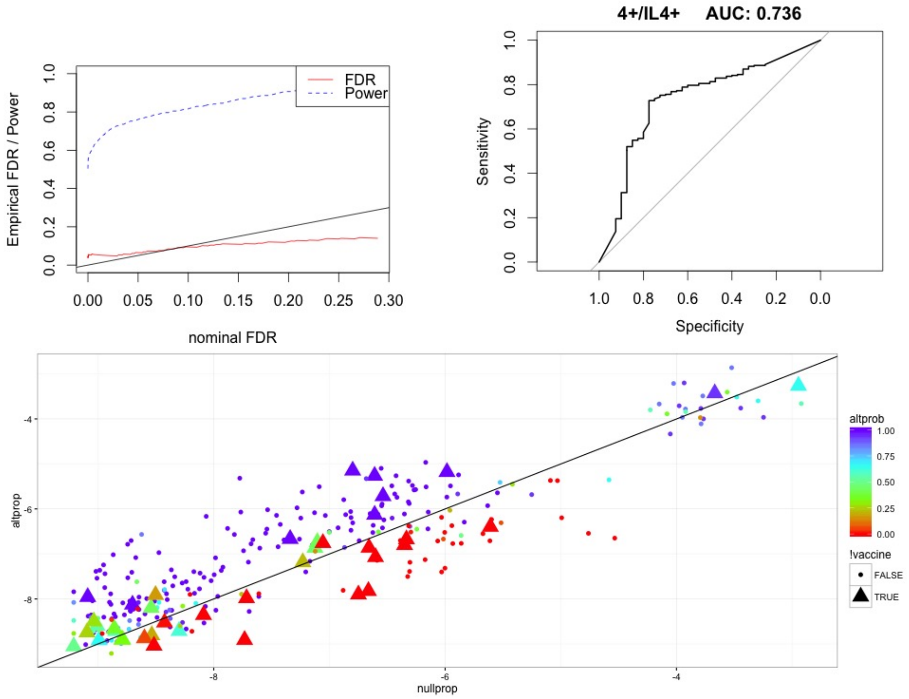

Figure 5: Posterior probabilities and classification errors for the joint model of T4+/CD154 and T4+/IL4+. The scatter plot is for log proportions of the T4+/IL4+ cell-subset.

2.2 Dependence Between Cell-Subsets

From the single random-effect analysis we learn that:

- It is beneficial to model multiple cell-subsets at the same time. Joint modeling can improve the subject analysis as well as (probably) the global effect estimates.
- Univariate random-effects are only suitable to modeling highly correlated cell subsets.

In order to better understand the dependence structure we can expect to see in this sort of datasets, we fit the following model to all pairs of T4+ cell types in the RV144 dataset. For the lth cell-subset, jth sample from the ith subject:

$$\text{logit}(p_{ijl}) = \beta_{0l} + \text{stim}_j * \beta_{1l} + \text{vaccine}_i * \beta_{2l} + \text{stim}_j * \text{vaccine}_i * \beta_{3l} + \nu_{il},$$

$$\nu_i \sim N_2(0, \Sigma).$$

The fitting procedure (as implemented in **lme4**) return the estimated covariance structure in each pairwise model. The resulting covariance and correlation structures are given in Tables 1 and 2 respectively. The highest correlations in the tables are between the four cell-subsets where there is a strong visible vaccine effect- CD154+, IFNg+, IL2+ and TNFa+. However, there are is some degree of non-negligible correlation between most cell-subsets.

	1	2	3	4	5	6	7
4+/CD154+	0.63	0.53	0.39	0.55	0.44	0.38	0.47
4+/IFNg+	0.53	0.69	0.44	0.31	0.29	0.20	0.53
4+/IL2+	0.39	0.44	0.39	0.13	0.23	0.11	0.47
4+/IL4+	0.55	0.31	0.13	3.13	0.95	0.61	0.18
4+/IL17a+	0.44	0.29	0.23	0.95	0.75	0.38	0.36
4+/MIP1B+	0.38	0.20	0.11	0.61	0.38	1.47	0.21
4+/TNFa+	0.47	0.53	0.47	0.18	0.36	0.21	0.68

Table 1: Covariance Between Random Effects in the RV144 Dataset.

	1	2	3	4	5	6	7
4+/CD154+	1.00	0.80	0.78	0.39	0.64	0.40	0.72
4+/IFNg+	0.80	1.00	0.85	0.21	0.40	0.20	0.77
4+/IL2+	0.78	0.85	1.00	0.12	0.42	0.15	0.92
4+/IL4+	0.39	0.21	0.12	1.00	0.62	0.29	0.13
4+/IL17a+	0.64	0.40	0.42	0.62	1.00	0.37	0.51
4+/MIP1B+	0.40	0.20	0.15	0.29	0.37	1.00	0.21
4+/TNFa+	0.72	0.77	0.92	0.13	0.51	0.21	1.00

Table 2: Correlations Between Random Effects in the RV144 Dataset.

3 Modeling with a Multivariate Random-Effect Distribution

Let y_{ijl} denote the count for the lth cell subset in the jth stimulation from the ith subject. We wish to estimate the model:

$$y_{ijl} \sim \pi_0 Bin(N_{ij}, p_{ijl0}) + \pi_1 Bin(N_{ij}, p_{ijl1})$$

$$\pi_0 + \pi_1 = 1,$$
$$\text{logit}(p_{ijlk}) = X_{ijl}\beta_l + T_{ijl}\tau_{lk} + \nu_{il},$$
$$\nu_i \sim N_L(0, \Sigma),$$
$$\tau_{kl} = \begin{cases} \tau_l & k = 1 \\ 0 & k = 0 \end{cases}.$$

Where X_{ijl} and T_{ijl} are known matrices, τ_{lk} is a random regression coefficient that equals either zero or an unknown constant (which needs to be estimated). The likelihood of the model is given by:

$$\mathcal{L}(\beta, \tau, \pi, \Sigma; y) = \prod_{i=1}^{n}\left[\sum_{k=0}^{1} \pi_k \int_{\mathbb{R}^L} \prod_{jl} f(y_{jl}|\tau_k, \beta, \nu_i)\varphi(\nu_i; \Sigma)d\nu_i\right].$$

3.1 Estimating the Mixed-Multivariate-Mixed-Effect-Model

The likelihood is non-trivial to maximize because it includes both a summation and a high-dimensional integral. In order to do so, we will use a nested Monte-Carlo EM algorithm where we have an outer MCEM algorithm, the M-step of which is itself an MCEM. Write the full-information log-likelihood for the mixture model:

$$l(\beta, \tau, \Sigma|\pi_{1,\ldots,n}) = \sum_{i=1}^{n}\sum_{k=0}^{1} \pi_{ik} \log \int_{\mathbb{R}^L} \prod_{jl} f(y_{jl}|\beta, \tau_k, \nu_i)\varphi(\nu_i; \Sigma)d\nu_i + \pi_{ik} \log \pi_k$$

$$:= \sum_{i=1}^{n}\sum_{k=0}^{1} \pi_{ik} \log f_k(y_i) + \pi_{ik} \log \pi_k.$$

3.1.1 Outer MCEM

Assuming we can compute $f_k(y_i)$, we get a straight forward MCEM procedure. Assume for now that the quantities $\hat{\mu}_{1k}, ...\hat{\mu}_{nk}$ and $\hat{\Sigma}$ are given.

- **E-Step**: For $i = 1, ...n$ and $k = 0, 1$ Sample

$$v_{1k}, ..., v_{Mk} \sim N(\mu_{ik}, \Sigma),$$

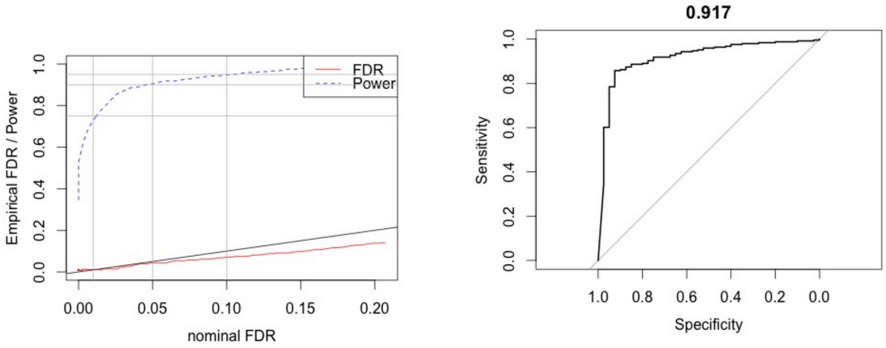

Figure 6: ROC and FDR for multivariate analysis of the RV144 dataset. In the left plot horizontal grey lines are at 0.75, 0.9 and 0.95.

and set:

$$w_{i0}^t = \left(1 + \frac{\pi_1}{\pi_0} \frac{\sum_{m=1}^M \frac{\varphi(v_{1m};0,\hat{\Sigma})}{\varphi(v_{1m};\hat{\mu}_{i1},\hat{\Sigma})} \prod_j f_1(y_{ij}; v_{m1})}{\sum_{m=1}^M \frac{\varphi(v_{0m};0,\hat{\Sigma})}{\varphi(v_{0m};\hat{\mu}_{i0},\hat{\Sigma})} \prod_j f_0(y_{ij}; v_{m0})}\right)^{-1}$$

$$\pi_{ik}^t = \frac{(t-1)\pi_{ik}^{t-1} + w_{ik}^t}{t}.$$

- **S-Step:** Sample $s_i^t \sim Ber(\pi_{ik}^t)$.
- **M-Step:** Set:

$$\hat{\pi}_k^t = \frac{\sum_{i=1}^n \pi_{ik}^t}{n},$$

$$\hat{\theta}^t := (\hat{\beta}, \hat{\tau}, \hat{\Sigma})^t = \arg\max \theta \sum_{i=1}^n \sum_{k=0}^1 \log f_k(y_i) I\{s_i^t = k\}.$$

3.1.2 Inner MCEM

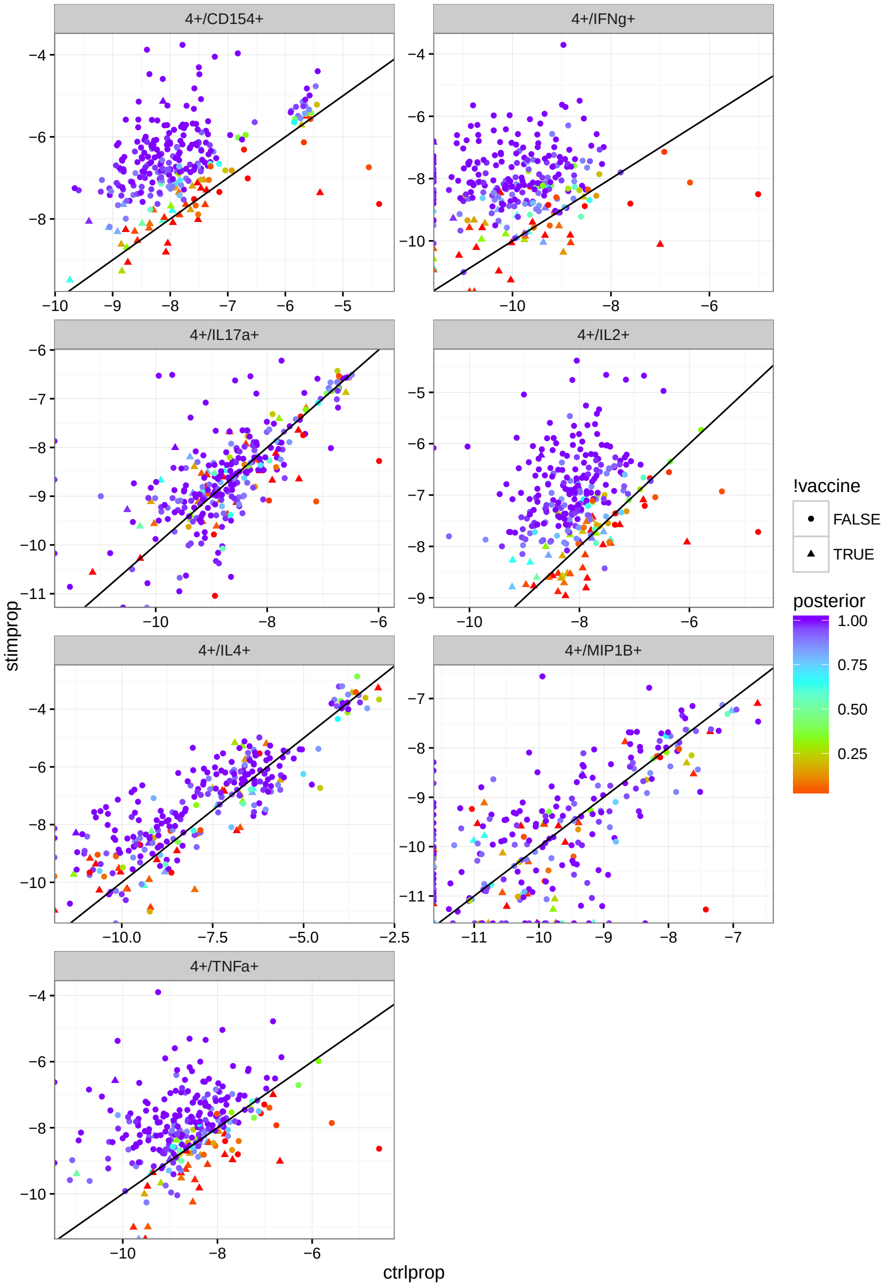

Figure 7: Scatter plot of log-proportions with posterior probabilities form multivariate fit.

	1	2	3	4	5	6	7
4+/CD154+	0.93	0.85	0.43	0.50	1.03	0.52	0.64
4+/IFNg+	0.85	1.07	0.29	0.57	0.80	0.35	0.71
4+/IL17a+	0.43	0.29	0.76	0.28	0.84	0.36	0.40
4+/IL2+	0.50	0.57	0.28	0.46	0.26	0.20	0.55
4+/IL4+	1.03	0.80	0.84	0.26	4.03	0.82	0.42
4+/MIP1B+	0.52	0.35	0.36	0.20	0.82	1.30	0.30
4+/TNFa+	0.64	0.71	0.40	0.55	0.42	0.30	0.79

Table 3: Covariance Between Random Effects in the RV144 Dataset - Multivariate Analysis.

	1	2	3	4	5	6	7
4+/CD154+	1.00	0.85	0.51	0.77	0.53	0.47	0.75
4+/IFNg+	0.85	1.00	0.33	0.81	0.38	0.30	0.78
4+/IL17a+	0.51	0.33	1.00	0.47	0.48	0.36	0.52
4+/IL2+	0.77	0.81	0.47	1.00	0.19	0.26	0.91
4+/IL4+	0.53	0.38	0.48	0.19	1.00	0.36	0.23
4+/MIP1B+	0.47	0.30	0.36	0.26	0.36	1.00	0.29
4+/TNFa+	0.75	0.78	0.52	0.91	0.23	0.29	1.00

Table 4: Correlations Between Random Effects in the RV144 Dataset - Multivariate Analysis.