Web-based Supplementary Materials for Combinatorial Mixture Models for Single-Cell Assays with Application to Vaccine Studies by Greg Finak, Steve De Rosa, Mario Roederer and Raphael Gottardo

Web Appendix A: Computational details for the beta-binomial model

Marginal likelihood derivations

For a given subject i, the null marginal likelihood is obtained after integrating out the prior from the likelihood for model $\mathcal{M}_0$, as follows,

$$\begin{aligned}
\mathrm{L}_0(\alpha_u, \beta_u | n_{si}, n_{ui}) &= \int_0^1 \Pr(n_{si}, n_{ui}|p_u)\pi(p_u|\alpha_u, \beta_u)\mathrm{d}p_u \\
&= \int_0^1 Pr(n_s|p_u)Pr(n_u|p_u)\pi(p_u|\alpha_u, \beta_u)\mathrm{d}p_u \\
&= \int_0^1 \binom{N_s}{n_s} p_u^{n_s}(1-p_u)^{N_s-n_s}\binom{N_u}{n_u} p_u^{n_u}(1-p_u)^{N_u-n_u}\frac{1}{\mathrm{B}(\alpha_u, \beta_u)}p_u^{\alpha_u-1}(1-p_u)^{\beta_u-1}\mathrm{d}p_u \\
&= \binom{N_s}{n_s}\binom{N_u}{n_u}\frac{1}{\mathrm{B}(\alpha_u, \beta_u)}\int_0^1 p_u^{n_s+n_u+\alpha_u-1}(1-p_u)^{N_s+N_u-n_s-n_u+\beta_u-1}\mathrm{d}p_u \\
&= \binom{N_s}{n_s}\binom{N_u}{n_u}\frac{\mathrm{B}(n_s+n_u+\alpha_u, N_s+N_u-n_s-n_u+\beta_0)}{\mathrm{B}(\alpha_u, \beta_u)}.
\end{aligned}$$

Similarly, the alternative marginal likelihood for a given subject is defined as,

$$\begin{aligned}
\mathrm{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s | n_{si}, n_{ui}) &= \int_0^1\int_0^1 \Pr(n_{si}, n_{ui}|p_u, p_s)\pi(p_u, p_s|\alpha_u, \beta_u, \alpha_s, \beta_s)\mathrm{d}p_s\mathrm{d}p_u \\
&= \int_0^1\int_0^1 \binom{N_s}{n_s} p_s^{n_s}(1-p_s)^{N_s-n_s}\binom{N_u}{n_u} p_u^{n_u}(1-p_u)^{N_u-n_u}\frac{1}{\mathrm{B}(\alpha_u, \beta_u)}p_u^{\alpha_u-1}(1-p_u)^{\beta_u-1}\cdot \\
&\quad \frac{1}{\mathrm{B}(\alpha_s, \beta_s)}p_s^{\alpha_s-1}(1-p_s)^{\beta_s-1}\mathrm{d}p_s\mathrm{d}p_u \\
&= \binom{N_s}{n_s}\binom{N_u}{n_u}\frac{1}{\mathrm{B}(\alpha_u, \beta_u)}\frac{1}{\mathrm{B}(\alpha_s, \beta_s)}\int_0^1 p_u^{n_u+\alpha_u-1}(1-p_u)^{N_u-n_u+\beta_u-1}\cdot \\
&\quad \int_0^1 p_s^{n_s+\alpha_s-1}(1-p_s)^{N_s-n_s+\beta_s-1}\mathrm{d}p_s\mathrm{d}p_u \\
&= \binom{N_s}{n_s}\binom{N_u}{n_u}\frac{\mathrm{B}(n_u+\alpha_u, N_u-n_u+\beta_u)\mathrm{B}(n_s+\alpha_s, N_s-n_s+\beta_s)}{\mathrm{B}(\alpha_u, \beta_u)\mathrm{B}(\alpha_s, \beta_s)}.
\end{aligned}$$

MCMC algorithm

In what follows, we use $(x|y)$ to denote the conditional distribution of x given y. In particular, we use $(x|\cdots)$ to denote the distribution of x conditional on everything else in the model. Our MCMC algorithms cycles through the following steps:

1. Update each $\alpha_u, \beta_u, \alpha_s$ and β_s by Metropolis-Hastings using a Gaussian symmetric proposal where the variance of the proposal is tuned for each parameter using the approach of Gelman et al. (2004); Raftery and Lewis (1992); Raftery (1996).

2. Update w by Gibbs sampling using the full conditional,

$$(w|\cdots) \sim \text{Beta}(\sum_i z_i, \sum_i (1 - z_i)).$$

3. for each i, update z_i by Gibbs sampling using the following full conditional,

$$(z_i|\cdots) \sim \text{B}(1, p_i).$$

where,

$$p_i = \frac{w \cdot \text{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s | n_{ui}, n_{si})}{w \cdot \text{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s | n_{ui}, n_{si}) + (1 - w) \cdot \text{L}_0(\alpha_u, \beta_u | n_{ui}, n_{si})}.$$

For each updated parameter, step 1 above involves the calculation of the following acceptance ratio, (e.g. α_u, below),

$$\frac{\text{L}(\alpha_u^{\text{new}}, \beta_u, \alpha_s, \beta_s | \cdots)\pi(\alpha_u^{\text{new}})}{\text{L}(\alpha_u, \beta_u, \alpha_s, \beta_s | \cdots)\pi(\alpha_u)}$$

where L is the complete marginal likelihood conditional on $\mathbf{z}$ defined as,

$$\text{L}(\alpha_u, \beta_u, \alpha_s, \beta_s | \mathbf{n}_u, \mathbf{n}_s, \mathbf{z}) = \prod_i \text{L}_{z_i}(\alpha_u, \beta_u, \alpha_s, \beta_s | \mathbf{n}_u, \mathbf{n}_s)$$

and π is the prior distribution of α_u. The obvious changes in the above expression are made for the acceptance ratios of $\alpha_s, \beta_s, \beta_u$. In our case each parameter has the same exponential prior with mean $1,000$

Web Appendix B: Constrained beta-binomial model

We can define a model where we constrain the stimulated proportions under the alternative model such that $p_s > p_u$. In this case, the only changes required are for the alternative marginal likelihood L_1 defined in the main manuscript by (1). Due to the constraint, the normalizing constant of the prior under the alternative (model $\mathcal{M}_1$) is not given by $\text{B}(\alpha_u, \beta_u)\text{B}(\alpha_s, \beta_s)$ but requires computing

$$Z(\alpha_u, \beta_u, \alpha_s, \beta_s) = \int_0^1 {p_u}^{\alpha_u - 1}(1 - p_u)^{\beta_u - 1} \int_{p_u}^1 p_s^{\alpha_s - 1}(1 - p_s)^{\beta_s - 1} dp_s dp_u.$$

Using this expression, the constrained alternative marginal likelihood can be written as

$$\text{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s | \mathbf{y}_i) = \binom{N_{ui}}{n_{ui}} \binom{N_{si}}{n_{si}} \frac{Z(n_{ui} + \alpha_u, N_{ui} - n_{ui} + \beta_u, n_{si} + \alpha_s, N_{si} - n_{si} + \beta_s)}{Z(\alpha_u, \beta_u, \alpha_s, \beta_s)}.$$

In general, there is no closed-form expression for $Z(\cdot)$, and a numerical approximation must be used. Let us denote by $\tilde{Z}(\alpha_u, \beta_u, \alpha_s, \beta_s)$ the approximation. A natural way to estimate $\tilde{Z}$ is to use Monte Carlo integration. Indeed, we can write

$$\tilde{Z}(\alpha_u, \beta_u, \alpha_s, \beta_s) = \text{B}(\alpha_u, \beta_u)\text{B}(\alpha_s, \beta_s) \int_0^1 \frac{{p_u}^{\alpha_u - 1}(1 - p_u)^{\beta_u - 1}}{\text{B}(\alpha_u, \beta_u)} (1 - F_{\alpha_s, \beta_s}(p_u)) dp_u \tag{1}$$

where F_{α_s, β_s} is the cumulative distribution function of a beta random variable with parameters α_s and β_s. Using this identity, it can be seen that $\tilde{Z}(\alpha_u, \beta_u, \alpha_s, \beta_s)$ can be approximated by

$$\tilde{Z}(\alpha_u, \beta_u, \alpha_s, \beta_s) \approx \text{B}(\alpha_u, \beta_u)\text{B}(\alpha_s, \beta_s) \sum_{k=1}^{K} (1 - F_{\alpha_s, \beta_s}(X_k))$$

where the X_k's are *iid* beta distributed random variables with parameters α_s, β_s and K is the number of terms used in the Monte Carlo approximation. This approximation works relatively well with our EM implementation and does not significantly increase the computing time. Unfortunately, the number of terms (*i.e.* value of K) required for the approximation to be good might be large and computing such a normalizing constant at each iteration would significantly slow down our MCMC implementation. As it tuns out, a better approximation can be obtained when α_s and β_s are integers. In this case, the cdf function in (1) can be calculated exactly using integration by parts, as follows,

$$F_{\alpha_s,\beta_s}(p_u) = \sum_{j=\beta_s}^{\beta_s+\alpha_s-1} \frac{(\beta_s+\alpha_s-1)!}{j!(\beta_s+\alpha_s-j)!}(1-p_u)^j p_u^{\beta_s+\alpha_s-j}.$$

Then using this identity, we obtain

$$\begin{aligned} Z(\alpha_u,\beta_u,\alpha_s,\beta_s) &= \mathrm{B}(\alpha_s,\beta_s) \sum_{j=\beta_s}^{\beta_s+\alpha_s-1} \frac{(\beta_s+\alpha_s-1)!}{j!(\beta_s+\alpha_s-j)!} \int_0^1 (1-p_u)^{\beta_u-1+j} p_u^{\alpha_u-1+\beta_s+\alpha_s-j} dp_u \\ &= \mathrm{B}(\alpha_s,\beta_s) \sum_{j=\beta_s}^{\beta_s+\alpha_s-1} \frac{(\beta_s+\alpha_s-1)!}{j!(\beta_s+\alpha_s-j)!} \mathrm{B}(\beta_u+j)\mathrm{B}(\alpha_u+\beta_s+\alpha_s-j). \end{aligned}$$

Typically, in ICS data α_s is relatively small leading to relatively few terms in the sum. However, the use of this exact identity in our MCMC algorithm requires the use of discrete priors on α_s and β_s, which can be restrictive in terms of fit (*e.g.*, if the true α_s is less than one) and can render mixing in the MCMC more difficult. In addition, even though the computation is exact and much faster for small values of α_s, which is typically the case with ICS data, it is still more demanding than the unconstrained model. In our case, we have decided to use the unconstrained model and simply fix the z_i to zero if the empirical proportion for the un-stimulated sample, p_u, is greater than that of the stimulation sample, p_s. Indeed, in the one-sided case, if $p_u > p_s$ the associated individual should be a non-responder and thus $z_i = 0$. In our experience, this computational shortcut performs just as well as the true one-sided implementation while being computationally much less demanding.

Web Appendix C: Computational details for the Dirichlet-multinomial model

Marginal likelihood derivations

Because our Dirichlet-multinomial is a direct extension of the beta-binomial model, the marginal likelihoods are obtained in the exact same fashion. For a given subject, the null marginal likelihood is defined as

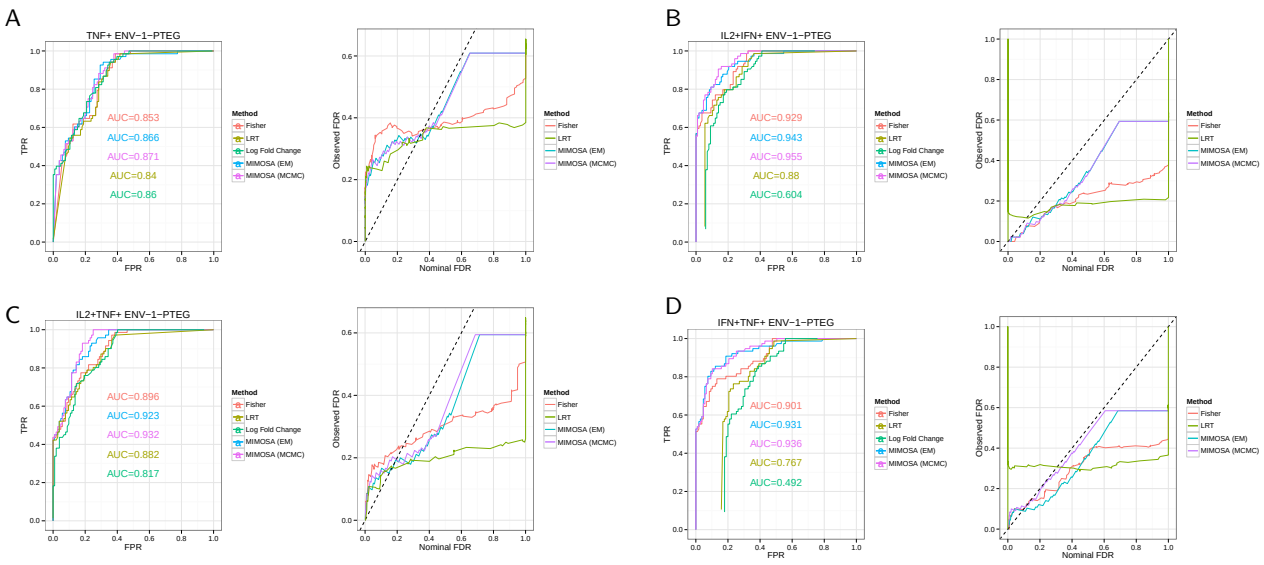
$$\begin{aligned} \mathrm{L}_0(\boldsymbol{\alpha}_u|\mathbf{n}_{si},\mathbf{n}_{ui}) &= \int\cdots\int \Pr(\mathbf{n}_{si},\mathbf{n}_{ui}|\mathbf{p}_u)\pi(\mathbf{p}_u|\boldsymbol{\alpha}_u)\mathrm{d}\mathbf{p}_u \\ &= \int\cdots\int \frac{\mathbf{N}_s!}{\prod_k n_{sik}!}\prod_k p_{uk}^{n_{sik}} \frac{\mathbf{N}_u!}{\prod_k n_{uik}!}\prod_k p_{uk}^{n_{uik}} \frac{1}{\mathrm{B}(\boldsymbol{\alpha}_u)}\prod_k p_{uk}^{\alpha_{uk}-1}\mathrm{d}\mathbf{p}_u \\ &= \frac{\mathbf{N}_s!}{\prod_k n_{sik}!}\frac{\mathbf{N}_u!}{\prod_k n_{uik}!}\frac{1}{\mathrm{B}(\boldsymbol{\alpha}_u)}\int_0^1 \prod_k p_{uk}^{n_{sik}+n_{uik}+\alpha_{uk}-1}\mathrm{d}\mathbf{p}_u \\ &= \frac{\mathbf{N}_s!}{\prod_k n_{sik}!}\frac{\mathbf{N}_u!}{\prod_k n_{uik}!}\frac{\mathrm{B}(\mathbf{n}_{si}+\mathbf{n}_u+\boldsymbol{\alpha}_u)}{\mathrm{B}(\boldsymbol{\alpha}_u)} \end{aligned}$$

Similarly, the alternative marginal likelihood for a given subject is:

$$\begin{aligned} \mathrm{L}_1(\boldsymbol{\alpha}_u,\boldsymbol{\alpha}_s|\mathbf{n}_{ui},\mathbf{n}_{si}) &= \frac{\mathbf{N}_s!}{\prod_k n_{sik}!}\frac{\mathbf{N}_u!}{\prod_k n_{uik}!}\frac{1}{\mathrm{B}(\boldsymbol{\alpha}_u)\mathrm{B}(\boldsymbol{\alpha}_s)}\int\cdots\int\prod_k p_{uk}^{n_{uik}+\alpha_{uk}-1}\int\cdots\int\prod_k p_{sk}^{n_{sik}+\alpha_{sk}-1}\mathrm{d}\mathbf{p}_u\mathrm{d}\mathbf{p}_s \\ &= \frac{\mathbf{N}_s!}{\prod_k n_{sik}!}\frac{\mathbf{N}_u!}{\prod_k n_{uik}!}\frac{\mathrm{B}(\mathbf{n}_{ui}+\boldsymbol{\alpha}_u)\mathrm{B}(\mathbf{n}_{si}+\boldsymbol{\alpha}_s)}{\mathrm{B}(\boldsymbol{\alpha}_u)\mathrm{B}(\boldsymbol{\alpha}_s)}. \end{aligned}$$

MCMC algorithm

The MCMC algorithm for the Dirichlet-multinomial model is analogous to the beta-binomial, above. The parameter vectors $\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_u$ are updated component-wise:



Web Figure A: Comparison of MIMOSA on other cytokines and cytokine combinations for ENV-1-PTEG stimulated CD4+ T-cells from the HVTN065 trial.

1. Update each α_{sk}, α_{uk} using a Gaussian symmetric proposal distribution with the variance of each proposal tuned using the approach of Gelman et al. (2004).

2. Update w by Gibbs sampling using the full conditional,

$$(w|\cdots) \sim \text{Beta}(\sum_i z_i, \sum_i (1 - z_i))$$

3. For each i, update z_i using the full conditional,

$$(z_i|\cdots) \sim \text{B}(1, p_i)$$

where

$$p_i = \frac{w \cdot \text{L}_1(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\mathbf{n}_{si}, \mathbf{n}_{ui})}{w \cdot \text{L}_1(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\mathbf{n}_{si}, \mathbf{n}_{ui}) + (1 - w) \cdot \text{L}_0(\boldsymbol{\alpha}_u|\mathbf{n}_{si}, \mathbf{n}_{ui})}$$

For each parameter component updated in step 1 above, compute the acceptance ratio (e.g. α_{uk}, below):

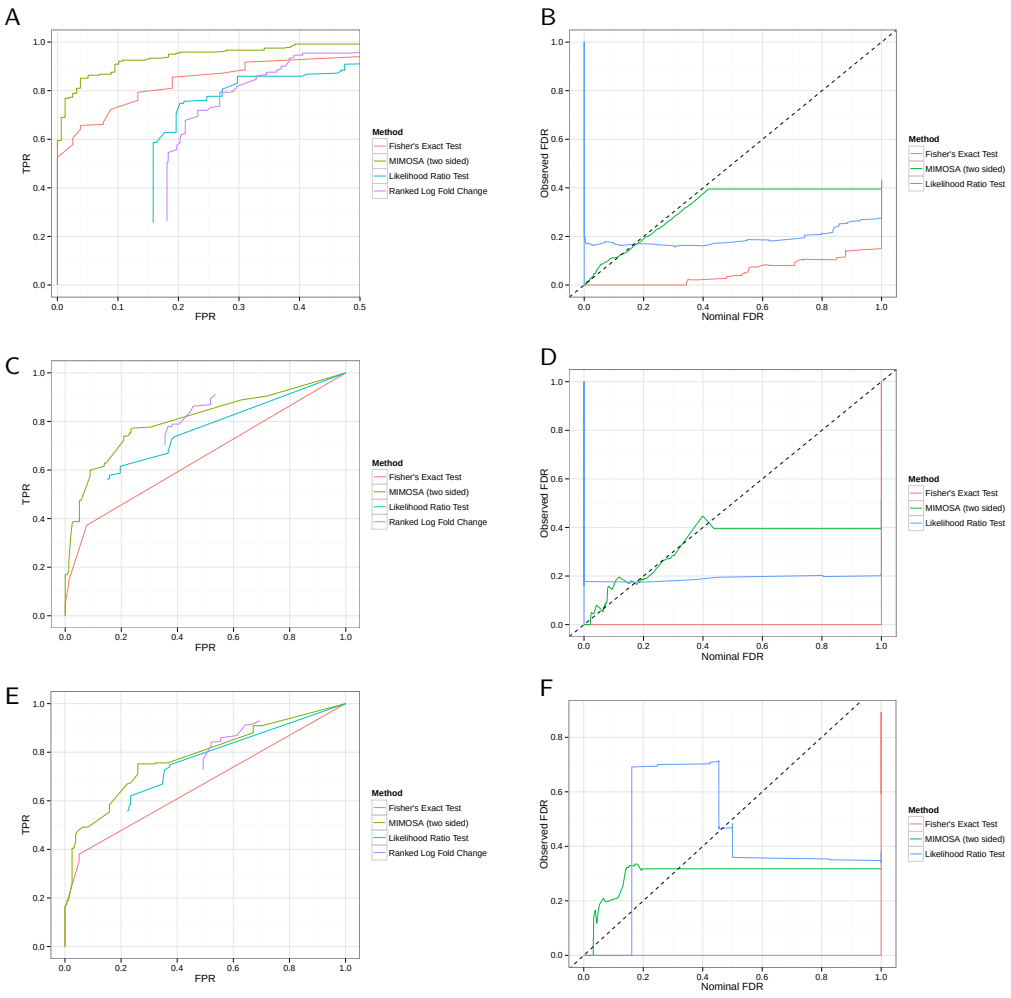
$$\frac{\text{L}(\alpha_{uk}^{\text{new}}, \boldsymbol{\alpha}_{u\{-k\}}, \boldsymbol{\alpha}_s|\cdots)\pi(\alpha_{uk}^{\text{new}})}{\text{L}(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\cdots)\pi(\alpha_{uk})}$$

where π is the prior distribution of the parameter, and $\boldsymbol{\alpha}_{u\{-k\}} = \{\alpha_{uj} : j \neq k\}$. Again, we have used the same exponential prior with mean 1,000 for each parameter. L is the complete marginal likelihood conditional on $\mathbf{z}$, defined as

$$\text{L}(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\mathbf{n}_{si}, \mathbf{n}_{ui}, \mathbf{z}) = \prod_i \text{L}_{z_i}(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\mathbf{n}_{si}, \mathbf{n}_{ui}).$$

References

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.



Web Figure B: Unconstrained MIMOSA model fit to data from a model violating model assumptions and to two-sided data with small counts. Data was simulated from a model where proportions were sampled from a truncated normal distribution over $[0, 1]$ rather than a Beta distribution. A) The average ROC from 10 simulation with N=5,000 events. B) The average observed and nominal FDR from 10 simulations with N=5,000 events. C) Average ROC for N=1,000 events D) Average observed and nominal FDR for N=1,000 events. MIMOSA fit to two-sided data were simulated from the standard model with N=1,000 events. E) Average ROC curves from 10 simulations and F) average observed vs nominal FDR from 10 simulations.

Raftery, A. (1996). Markov Chain Monte Carlo in Practice - Walter R. Gilks, Sylvia Richardson, D. J. Spiegelhalter. *Markov chain Monte Carlo in practice* .

Raftery, A. E. and Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *STATISTICAL SCIENCE* **7,** 493–497.