SUPPLEMENTARY METHODS

Baysian analysis of metabolite profiles

We consider data in which substances of different known (geno-)types $t = 1, \ldots, T$ are processed under slightly different experimental conditions $c = 1, \ldots, C$. There are R_{ct} replicates of type t under condition c. The data for each replicate consist of observations y_{ctmr}, where the index for metabolite varies in the range $m = 1, \ldots, M$.

We suppose that type-metabolite combinations are switched on or off independently with probability p, and that $\gamma_{tm} = 1$ or $\gamma_{tm} = 0$ respectively. If a combination is switched off, then the profile mean for that metabolite-type pair takes a baseline value μ, while if it is switched on, the profile has a normal distribution, as outlined below.

The proposed model is the hierarchy

$$\begin{aligned} y_{ctmr} \mid \eta_{ctm}, \sigma^2 &\overset{\text{iid}}{\sim} N(\eta_{ctm}, \sigma^2), \\ \eta_{ctm} \mid \theta_{tm}, \tau_\eta^2 &\overset{\text{iid}}{\sim} N(\theta_{tm}, \tau_\eta^2), \\ \theta_{tm} \mid \gamma_{tm} &\overset{\text{iid}}{\sim} N(\mu, \gamma_{tm}\tau_\theta^2), \\ \gamma_{tm} &\overset{\text{iid}}{\sim} B(p), \end{aligned}$$

where $N(a, b)$ denotes the normal distribution with mean a and variance $b > 0$, and $B(p)$ denotes the Bernoulli distribution with probability $p \in (0, 1)$. The parameters γ_{tm} indicate whether or not each metabolite-type combination is on. If so, then the profile mean under ideal experimental conditions is θ_{tm}, which is taken to have a normal distribution. However the actual experimental conditions will not be ideal, and this will induce a difference between the ideal profile and that actually observeable. Finally the value observed y_{ctmr} is normally distributed around the observeable profile. If $\gamma_{tm} = 0$, then the same hierarchy applies, but with $\theta_{tm} = \mu$. Thus the model can be written as

$$y_{ctmr} = \mu + \gamma_{tm}(\theta_{tm} - \mu) + (\eta_{ctm} - \gamma_{tm}\theta_{tm}) + \varepsilon_{ctmr},$$

where the terms in parentheses represent the effects first of the 'switching on' of a metabolite-type combination, and the second represents the effect of the difference between ideal and actual experimental conditions.

Under this mixture model the marginal distribution of y_{ctmr} may be written as

$$(1 - p)N(\mu, \tau_\eta^2 + \sigma^2) + pN(\mu, \tau_\theta^2 + \tau_\eta^2 + \sigma^2), \tag{1}$$

and the joint density of the replicates y_{ctmr}, $r = 1, \ldots, R_{ctm}$ for a given combination of condition, type, and metabolite is readily obtained. It is then straightforward to obtain the joint density of the observations and hence to produce empirical Bayes estimates of the five parameters $\mu, p, \sigma^2, \tau_\eta^2, \tau_\theta^2$. Notice that data for the same metabolite-type combination under different experimental conditions are not in fact needed—intuitively because if $R_{ctm} > 1$, then σ^2 may be estimated from the sum of squares for error, and then τ_η^2 and τ_θ^2 are identifiable from the mixture (1). In fact maximum likelihood estimation is applied, and yields $\widehat{p} \doteq 0.03$, $\widehat{\mu} \doteq 0.05$, $\widehat{\sigma}^2 = 0.18$, $\widehat{\tau}_\eta^2 \doteq 2.2$, $\widehat{\tau}_\theta^2 \doteq 8$ for the data described in the paper. More precise estimates of the variance components could be obtained by including the data on known deg mutants in the likelihood, assuming that they are the same genotype but reared under different experimental conditions.

Once the hyperparameters are estimated a Bayesian classification rule can be applied, as follows. Let y_{tm} denote the data available for a given metabolite m and known type t, and suppose the corresponding data of unknown type are z_m. Let y and z denote the totality of the known data and the data to be classified, respectively, and let $U = 1, \ldots, T$ denote the unknown type of the data z. To assign z to one of the known types, we compute the posterior probabilities

$$\Pr(U = u \mid z, y), \quad u = 1, \ldots, T,$$

which equal

$$\frac{f(z, y \mid U = u)\Pr(U = u)}{\sum_{u'=1}^{T} f(z, y \mid U = u')\Pr(U = u')}, \quad u = 1, \ldots, T. \tag{2}$$

We set the prior probabilities $\Pr(U = u)$ to be equal, so for the (m, t) metabolite-type combination we need to compute

$$\begin{aligned} f(z_m, y_{tm} \mid U = u) &= (1-p) f(z_m, y_{um} \mid \gamma_{tm} = 0) \prod_{t \neq u, m \neq n} f(y_{tm} \mid \gamma_{tm} = 0) + \\ & \qquad p f(z_m, y_{um} \mid \gamma_{tm} = 1) \prod_{t \neq u, m \neq n} f(y_{tm} \mid \gamma_{tm} = 1). \end{aligned}$$

This can be rewritten as

$$f(z_m, y_{um} \mid U = u) = (1-p) A_m f(z_m \mid y_{um}, \gamma_{um} = 0) + p B_m f(z_m \mid y_{um}, \gamma_{um} = 1)$$

where

$$A_m = \prod_t f(y_{tm} \mid \gamma_{um} = 0), \quad B_m = \prod_t f(y_{tm} \mid \gamma_{um} = 1)$$

are fixed for the mth metabolite.

This computation applied to the data yields the probabilities given in the paper.

We measure the importance of a metabolite-type combination using the log Bayes factor for comparison of $\gamma_{tm=1}$ and $\gamma_{tm} = 0$, that is,

$$2 \log B_{10} = \frac{B_m}{A_m};$$

this is also called the weight of evidence. We use the usual (rather crude) scale for this: $0 < 2\log B_{10} \leq 2$ (low); $2 < 2\log B_{10} \leq 6$ (medium); $6 < 2\log B_{10} \leq 10$ (high); $10 < 2\log B_{10}$ (very strong).