Supporting Information
Practical considerations

Rmpi

metaMix uses the Rmpi package \( \text{http://cran.r-project.org/web/packages/Rmpi/} \). Rmpi is the R wrapper to MPI, the Message-Passing Interface. MPI is the standardized and portable system allowing parallel programs communicate with each other.

Number of chains and tempering scheme

The optimal number of chains used in the parallel tempering is not obvious. The main idea is that the number of chains used must be large enough to ensure successful swaps between all neighboring chains. The first limitation is the number of chains we can run on a computer. The computing facility we use for all analyses is the UCL CS cluster. The availability of suitable machines as well as considerations towards minimizing the queuing time of the submitted jobs, defines the maximum number we can use as \( N = 12 \) chains.

The choice of temperature values is motivated by the fact that these must not be too far apart, so that exchange of values between the chains can occur. Additionally the maximum value must be high enough so that no chains become trapped in local minima, hence allowing for global moves.

We implemented a power decay heating scheme:

\[
t_n = (t_{n-1} - K)^\alpha, \text{ where } n = 2, \ldots, N, \ K \in (0, 1) \text{ and } \alpha > 1
\]

and using \( K = 0.001 \) and \( \alpha = 3/2 \) we achieve a slowly heating sequence of chains with a lot of chains similar to the target.

We find that for \( N < 10 \) the maximum temperature is not very high, hindering a quick global exploration. Ideally we would prefer to run 14 to 20 chains but given our computing constraint \( N = 12 \) performs satisfactorily.

Number of iterations

Given the described setup of our Parallel Tempering, we find that the MCMC produces reasonable results in \( (5 \times \text{number of potential species}) \) iterations for each chain.

Relative abundance estimation

We estimate the relative abundances \( w \) for a given set of species, using either a frequentist - the Expectation-Maximization algorithm - or a Bayesian approach - the Gibbs Sampler.

Let us recall the metagenomic mixture model notation introduced in the main text. The data consist of \( N \) sequencing reads \( X = (x_1, \ldots, x_N) \). The relative abundance proportions, that is the proportion of each of the \( k \) species
in the mixture is denoted by $w = (w_1, ..., w_K)$. The probability of observing the read $x_i$ conditional on the assumption that it originated from species $S_j$ is $p(x_i|x_i \text{ from } S_j) = p_{ij}$.

In the mixture model setting each read $x_i$, $1 \leq i \leq n$ is assumed to arise from a specific but unknown component of the mixture. The mixture structure is deconvoluted by the introduction of latent variables: we associate $x_i$ with $z_i = (z_{i1}, ..., z_{ik})$, a $k$-dimensional vector indicating to which component $x_i$ belongs.

$z_{ij} = \begin{cases} 
1 & \text{if } x_i \text{ belongs to class } j \\
0 & \text{otherwise} 
\end{cases}$

(2)

**Expectation Maximization approach**

The EM approach to parameter estimation obtains point estimates of the parameters by maximizing the likelihood [Dempster and Laird, 1977].

The algorithm iterations consist of two steps. In the first step, the expected value of the missing variables $z_i$ is computed based on $p(z|x, w)$. In the next step we calculate the new mixing parameters $w$ that maximize the expected complete-data log likelihood

$$
\mathbb{E}[\ln p(X, z|w)] = \sum_z p(z|x, w) \ln p(x, z|w)
$$

(3)

with the complete-data likelihood given by

$$
p(X, z|w) = \prod_{i=1}^{n} \prod_{j=1}^{k} (w_j p_{ij})^{z_{ij}}
$$

(4)

**EM algorithm**

- Initialization $w^{(0)}$
- At iteration $t$

1. **Expectation step.**
   Generate $z_i^{(t)}$ from $p(z_i^{(t)} = j|x_i, w_j^{(t-1)})$.

   $$
z_{ij} = \frac{p(z_i = j, x_i|w)}{p(x_i|w)} = \frac{p(z_i = j|w)p(x_i|z_i = j, w)}{\sum_{j=1}^{k} p(z_i = j|w)p(x_i|z_i = j, w)} = \frac{w_j p_{ij}}{\sum_{j=1}^{k} w_j p_{ij}}
$$

(5)

2. **Maximization step.**
   Given $z_i$ from E-step, calculate new $w^{(t)}$ that maximize the expectation of the complete-data log-likelihood (eq.3).
\[ w^{(t)} = \arg \max_w \mathbb{E}[\ln p(X, z|w)] \]  

(6)

It can be shown that this

\[ w^{(t)} = \frac{\sum_{i=1}^n z_i^{(t)}}{N} \]  

(7)

Gibbs sampling approach

The Gibbs sampler is a Markov Chain Monte Carlo method based on the successive simulation of \( z \) and \( w \) [Diebolt and Robert, 1994]. After convergence we obtain the full posterior distribution of \( w \).

Gibbs sampler

1. Initialization \( w^{(0)} \)

2. At iteration \( t \)
   1. Generate \( z_i^{(t)} \) from \( p(z_i^{(t)} = j|x_i, w^{(t-1)}_j) \). So
      \[ z_i \sim \text{Mult}(1; \hat{z}_i^{(t-1)}, \ldots, \hat{z}_{ik}^{(t-1)}) \]  
      (8)
      where \( \hat{z}_{ij} \) is given by equation [5]
   2. Compute \( n_j^{(t)} = \sum_{i=1}^n z_i^{(t)} \).
   3. Generate \( w^{(t)} \) from
      \[ \pi(w|z^{(t)}) \sim D(a_1 + n_1^{(t)}, \ldots, a_k + n_k^{(t)}) \]  
      (9)

[9] can be explained as \( \pi(w|z) \propto \pi(z|w)\pi(w) \), where for the mixing parameters \( w \) a conjugate prior \( \pi(w) \) is the Dirichlet distribution with parameters \( \alpha = \{\alpha_1, \ldots, \alpha_k\} \). Additionally:

\[ \pi(z|w) = \prod_{i=1}^n \pi(z_i|w) = \prod_i w_1^{z_i} \ldots w_k^{z_{ik}} = \prod_{i=1}^n \prod_{j=1}^k w_j^{z_{ij}} = \prod_{j=1}^k w_j^{n_j} \]  

(10)

hence sampling \( w \) from \( Dir \sim (\alpha_1 + n_1, \ldots, \alpha_k + n_k) \).
Importance Sampling, Defensive Sampling, MLE approximation

Accounting for the uncertainty in mixture weights $w$, a Monte Carlo approximation can be used for the marginal likelihood, by drawing independent samples from the prior to estimate $P(X)$ and averaging the likelihood. The simulation from the prior is computationally inefficient, as the majority of samples are outside the regions of high likelihood. Importance sampling (IS) techniques can be used to reduce the variance of the estimator. To obtain an efficient IS proposal distribution, we first approximate the posterior distribution $P(w|X)$ with a normal multivariate distribution $g$, based on the samples generated by the Gibbs sampler. We then generate $n$ samples $y_i \sim \mathcal{N}(\mu, \Sigma)$, $1 \leq i \leq n$. Defining the IS weights $a_i = \frac{\pi(y_i)}{g(y_i)}$, the IS estimator of the marginal likelihood is then:

$$ \hat{I} = \frac{\sum_i a_i P(X|y)}{\sum_i a_i} \tag{11} $$

The method works well if the importance distribution is fairly similar to the target distribution but with heavier tails. Using the posterior approximation as the importance distribution means that the IS estimator becomes the harmonic mean estimator (HME). HME is known to be unstable and to overestimate the marginal likelihood $P(X)$. In order to overcome this issue and to make the distribution tails heavier, we perform defensive importance sampling by using a mixture of posterior and prior as the IS distribution [Hesterberg, 1995]. This approach is only slightly costlier in computational time compared to the typical IS.

As discussed in the main text, for the IS we have approximated the posterior distribution $P(w|X)$ with a normal multivariate distribution $g$, based on the output of the Gibbs sampler. However the posterior approximation $g$ as IS proposal distribution means that the marginal likelihood estimator becomes the Harmonic Mean estimator. This estimator can be unstable; a solution to this problem is the defensive importance sampling [Hesterberg, 1995]. The main idea is the incorporation of a heavy tail component in the importance function $g$, effectively substituting it by the mixture density:

$$ \lambda g(y) + (1 - \lambda)q(y), \ 0 < \lambda < 1 \tag{12} $$

where $\lambda$ is close to 1. A natural choice for $q$ as the stabilizing factor is the prior density $\pi$.

In practice that means that the samples we use for the defensive IS estimator are generated with probability $\lambda$ from $g$ and with probability $1 - \lambda$ from $\pi$.

Comparison of the 3 approaches

We compared the performance of metaMix on the same simHC FAMEs dataset, using Importance Sampling and Defensive Importance Sampling (95% samples produced by posterior approximating $g$ and 5% by $\pi$) for the marginal likelihood
estimation as well as using the MLE approximation. The latter are the results presented in the main text. For the IS and the defensive IS at each MCMC iteration 1,000 samples were drawn from the proposal distribution.

The resulting species profiles can be seen in table S3. We ran the MCMC for 1,000 iterations in order to obtain results within 24 hours. All three approaches produced almost identical results, in terms of species identified and abundance estimates, with the defensive IS performing a bit better in terms of abundance estimation accuracy. However the MLE approximation method was ∼13x times faster than the other two, reducing the time required from ∼19 hours to 90 minutes.

Table 1: FAMeS simHC - comparing the effect of different marginal likelihood estimation methods on metaMix performance: species profiling, accuracy of abundance estimation and computational time.

<table>
<thead>
<tr>
<th></th>
<th>Importance Sampling</th>
<th>Defensive Sampling</th>
<th>MLE approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of species</td>
<td>116</td>
<td>116</td>
<td>116</td>
</tr>
<tr>
<td>( w ) estimate - rRMSE</td>
<td>17</td>
<td>16.8</td>
<td>17.1</td>
</tr>
<tr>
<td>( w ) estimate - AVGRE</td>
<td>8.5</td>
<td>8.3</td>
<td>8.6</td>
</tr>
<tr>
<td>Computational time (hours)</td>
<td>18.6</td>
<td>18.6</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Reference Database for FAMeS datasets

As mentioned in the main text, all NCBI bacterial genomes [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz) were used to construct the database. Additionally, there were six organisms in the dataset that were missing from the database, so we added these manually (taxon identifiers: 155919, 155920, 165597, 286604, 321955, 332415).
References

