Supplementary Material

A probabilistic model to recover individual genomes from metagenomes

Johannes Dröge[1], Alexander Schönhuth[2], and Alice C. McHardy[3]

[1] *Helmholtz Centre for Infection Research, Braunschweig, science@fungs.de*
[2] *Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, a.schoenhuth@cwi.nl*
[3] *Helmholtz Centre for Infection Research, Braunschweig, alice.mchardy@helmholtz-hzi.de*

2016-12-16

Supporting Data

The simulated contigs, features files and scripts to reproduce the results are deposited under:

DOI:10.5281/zenodo.201076

Supplementary Methods

Poisson approximation for absolute abundance

When sequencing reads have been mapped to the contigs, we can quantify the number of reads that covers each position of each contig. This is the vector x with $len(\boldsymbol{x}) = L$. We model the positional read coverage using a Poisson event model and assume that the positions are independent according to the Lander-Waterman statistics so that the joint likelihood is a product of positional likelihoods. Additionally, we scale the likelihood to a single event by taking the geometric mean. After simplification, the formula almost looks like the the Poisson over the mean contig coverage.

$$\mathcal{L}(\theta \mid \boldsymbol{x}) = \sqrt[L]{\prod_{i=1}^{L} \frac{\theta^{x_i}}{x_i!} e^{-\theta}} = \left(\frac{\prod_{i=1}^{L} \theta^{x_i}}{\prod_{i=1}^{L} x_i!} e^{-\theta L} \right)^{\frac{1}{L}} = \frac{\bar{\theta}}{\sqrt[L]{\prod_{i=1}^{L} x_i!}} e^{-\theta} \tag{1}$$

The data term in the denominator is a constant factor which is not dependent on θ. It is the geometric mean over the $x_i!$ values which we approximate using the arithmetic mean $\bar{x}$ of the positional contig coverage values.

$$\sqrt[L]{\prod_{i=1}^{L} x_i!} \approx \left(\frac{1}{L} \sum_{i=1}^{L} x_i \right)! = \bar{x}! \tag{2}$$

The approximation is good if the variance of the x_i is low. We use the approximation to avoid to handle other values than the mean which is usually computed. Since the term is a data constant, it is irrelevant for model comparison where only θ differs among the genomes. The approximated likelihood using mean values is the standard Poisson formula.

$$\mathcal{L}'(\theta \mid \boldsymbol{x}) = \frac{\theta^{\bar{x}}}{\bar{x}!} e^{-\theta} \tag{3}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell'(\theta \mid \boldsymbol{x}) = -\log \bar{x}! + \bar{x} \log \theta - \theta \tag{4}$$

MLE for Poisson

The multi-sample log-likelihood is the weighted sum over the sample log-likelihoods using mean vector $\boldsymbol{a_i}$ with length $len(\boldsymbol{a_i}) = M$. This corresponds to the geometric mean in the exponential likelihood formula.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{a_i}) = \frac{1}{M} \sum_{j=1}^{M} -\log a_{i,j}! + a_{i,j} \cdot \log \theta_j - \theta_j \tag{5}$$

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ on the training data a. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{a_i}) = \sum_{i=1}^{N} w_i \cdot \frac{1}{M} \sum_{j=1}^{M} -\log a_{i,j}! + a_{i,j} \log \theta_j - \theta_j \tag{6}$$

The partial derivative of f with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial f}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{a_{i,j}}{\theta_j} - 1 \right) \tag{7}$$

We find the zeros of f to determine the MLE $\hat{\theta}_j$.

$$\sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{a_{i,j}}{\theta_j} - 1 \right) = 0 \Leftrightarrow \sum_{i=1}^{N} \frac{w_i \, a_{i,j}}{\theta_j} = \sum_{i=1}^{N} w_i \Leftrightarrow \theta_j = \frac{\sum\limits_{i=1}^{N} w_i \, a_{i,j}}{\sum\limits_{i=1}^{N} w_i} \tag{8}$$

We see that the estimates for θ_j maximize the joint log-likelihood because the second partial derivative with respect to θ_j is always negative.

$$\frac{\partial^2 f}{\partial \theta_j^2} = -\sum_{i=1}^{N} \frac{w_i \, a_{i,j}}{M \, \theta_j^2} \tag{9}$$

Binomial approximation for relative abundance

Similarly to the Poisson approximation for absolute abundance, we derive the Binomial approximation via a product of positional Binomials. Vector $\boldsymbol{x}$ with length $len(\boldsymbol{x}) = L$ holds the positional read coverage of a contig with length L for one sample and vector s with same length holds the sum of positional read counts for the position i of the contig across all samples. There must be more than one sample to apply this model. We write the likelihood normalized to a single event as

$$\begin{aligned}
\mathcal{L}(\theta \mid \boldsymbol{x}) &= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i} \theta^{x_i} (1-\theta)^{(s_i - x_i)}} \\
&= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} \cdot \sqrt[L]{\prod_{i=1}^{L} \theta^{x_i}} \cdot \sqrt[L]{\prod_{i=1}^{L} (1-\theta)^{(s_i - x_i)}} \\
&= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} \cdot \theta^{\bar{x}} \cdot (1-\theta)^{(\bar{s} - \bar{x})}
\end{aligned} \tag{10}$$

The geometric mean of positional binomial coefficients (first term) is again a constant factor which is not dependent on θ. We approximate this term using the arithmetic mean.

$$\begin{aligned}
\sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} &= \frac{\sqrt[L]{\prod_{i=1}^{L} s_i!}}{\sqrt[L]{\prod_{i=1}^{L} x_i!} \cdot \sqrt[L]{\prod_{i=1}^{L} (s_i - x_i)!}} \\
&\approx \frac{\frac{1}{L}\sum_{i=1}^{L} s_i!}{\frac{1}{L}\sum_{i=1}^{L} x_i! \cdot \frac{1}{L}\sum_{i=1}^{L} (s_i - x_i)!} \\
&\approx \frac{\frac{1}{L}\sum_{i=1}^{L} s_i!}{\frac{1}{L}\sum_{i=1}^{L} x_i! \cdot \left(\frac{1}{L}\sum_{i=1}^{L} s_i - \frac{1}{L}\sum_{i=1}^{L} x_i\right)!} = \binom{\bar{s}}{\bar{x}}
\end{aligned} \tag{11}$$

The approximation is good if the differences in the coefficients are small. We use the approximation to avoid to handle other values than the mean which is usually computed. Since the term is a data constant, it is irrelevant for model comparison where only θ differs among the genomes. The approximated likelihood using mean values is the standard Binomial formula.

$$\mathcal{L}'(\theta \mid \boldsymbol{x}) = \binom{\bar{s}}{\bar{x}} \theta^{\bar{x}} (1-\theta)^{(\bar{s} - \bar{x})} \tag{12}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell'(\theta \mid \boldsymbol{x}) = \log \binom{\bar{s}}{\bar{x}} + \bar{x} \log \theta + (\bar{s} - \bar{x}) \log(1-\theta) \tag{13}$$

MLE for Binomial

The multi-sample log-likelihood is the weighted sum over the sample log-likelihoods using mean vector $\boldsymbol{r_i}$ with length $len(\boldsymbol{r_i}) = M$. This corresponds to the geometric mean in the exponential likelihood formula.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{r_i}) = \frac{1}{M} \sum_{j=1}^{M} \log \binom{R_i}{r_{i,j}} + r_{i,j} \log \theta_j + (R_i - r_{i,j}) \log(1 - \theta_j) \tag{14}$$

R_i is the sum of the abundance vector $\boldsymbol{r_i}$.

$$R_i = \sum_{j=1}^{M} r_{i,j} \tag{15}$$

Because both R_i and $r_{i,j}$ can be real numbers, we need to generalize the binomial coefficient to positive real numbers via the gamma function Γ.

$$\log \binom{n}{k} = \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \tag{16}$$

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ of the training data r. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$\begin{aligned} f(\boldsymbol{\theta}) &= \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{r_i}) \\ &= \sum_{i=1}^{N} w_i \cdot \frac{1}{M} \sum_{j=1}^{M} \log \binom{R_i}{r_{i,j}} + r_{i,j} \log \theta_j + (R_i - r_{i,j}) \log(1 - \theta_j) \end{aligned} \tag{17}$$

The partial derivative of f with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial f}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{r_{i,j}}{\theta_j} - \frac{R_i - r_{i,j}}{1 - \theta_j} \right) \tag{18}$$

We find the zeros of f to determine the MLE $\hat{\theta}_j$.

$$\begin{aligned} & \sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{r_{i,j}}{\theta_j} - \frac{R_i - r_{i,j}}{1 - \theta_j} \right) = 0 \\ \Leftrightarrow\, & (1 - \theta_j) \sum_{i=1}^{N} w_i r_{i,j} = \theta_j \left(\sum_{i=1}^{N} w_i R_i - \sum_{i=1}^{N} w_i r_{i,j} \right) \\ \Leftrightarrow\, & \frac{1}{\theta_j} \sum_{i=1}^{N} w_i r_{i,j} = \sum_{i=1}^{N} w_i R_i \\ \Leftrightarrow\, & \theta_j = \frac{\sum_{i=1}^{N} w_i r_{i,j}}{\sum_{i=1}^{N} w_i R_i} \end{aligned} \tag{19}$$

We see that the estimates for θ_j maximize the joint log-likelihood because the second partial derivative with respect to θ_j is negative for our estimates $\hat{\theta}_j$ for all $j \in \{1...M\}$.

$$\frac{\partial^2 f}{\partial \theta_j^2} = -\frac{R_i\theta_j^2 - 2r_{i,j}\theta_j + r_{i,j}}{(\theta_j - 1)^2\,\theta_j^2} \tag{20}$$

$$-\frac{R_i\hat{\theta}_j^2 - 2r_{i,j}\hat{\theta}_j + r_{i,j}}{\left(\hat{\theta}_j - 1\right)^2\hat{\theta}_j^2} < 0 \Leftrightarrow \sum_{i=1}^{N} w_i r_{i,j} < \sum_{i=1}^{N} w_i R_i \tag{21}$$

The last inequality is true by definition of R_i (assuming $r_{i,j} \neq R_i$ for simplicity).

Naïve Bayes model for nucleotide composition

The Naïve Bayes model assumes independence of features so that the likelihood can be written as a product of likelihoods for all features. The feature vector $\boldsymbol{x}$ for a contig contains nucleotide features such as all the absolute counts for all possible 5-mers. The length $len(\boldsymbol{x})$ is M. The total sum of counts for the contig is S.

$$S = \sum_{i=1}^{M} x_i \tag{22}$$

The likelihood is normalized to a single event via the geometric mean.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sqrt[S]{\prod_{i=1}^{M} \theta_i^{x_i}} = \prod_{i=1}^{M} \theta_i^{\frac{x_i}{S}} = \prod_{i=1}^{M} \theta_i^{x_i'} \tag{23}$$

Therefore, we directly use the normalized features.

$$x_i' = \frac{x_i}{\sum_{j=1}^{M} x_j} \tag{24}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x'}) = \sum_{i=1}^{M} x_i' \log \theta_i \tag{25}$$

MLE for Naive Bayes

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ on the training data c. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{c_i}) = \sum_{i=1}^{N} w_i \cdot \sum_{j=1}^{M} c_{i,j} \log \theta_j \tag{26}$$

We consider the constraint that $sum(\boldsymbol{\theta}) = 1$ because these are relative frequencies in each genome.

$$\sum_{j=1}^{M} \theta_j = 1 \tag{27}$$

Using the Lagrange method, we set up a function to maximize the joint data log-likelihood $f(\boldsymbol{\theta})$ under the given constraint.

$$\Lambda(\boldsymbol{\theta}, \lambda) = f(\boldsymbol{\theta}) + \lambda \left(\left(\sum_{j=1}^{M} \theta_j \right) - 1 \right) \tag{28}$$

The partial derivative of Λ with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial \Lambda}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i \, c_{i,j}}{\theta_j} + \lambda \tag{29}$$

We find the zeros of Λ to determine the MLE $\hat{\theta}_j$.

$$\frac{\partial \Lambda}{\partial \theta_j} = 0 \Leftrightarrow \theta_j = \frac{\sum_{i=1}^{N} w_i \, c_{i,j}}{-\lambda} \tag{30}$$

Substituting θ_j in Suppl. Equation 27 gives

$$-\lambda = \sum_{i=1}^{N} w_i \sum_{j=1}^{M} c_{i,j} = \sum_{i=1}^{N} w_i \tag{31}$$

The last simplification works because we work with normalized features that sum to one. Finally, we substitute $-\lambda$ in (1) for the MLE.

$$\hat{\theta}_j = \frac{\sum_{i=1}^{N} w_i \, c_{i,j}}{\sum_{i=1}^{N} w_i} \tag{32}$$

Hierarchic Naive Bayes model for sequence similarity

We adapted the Naive Bayes model to weighted taxa by transforming the associated weights (i.e. alignments scores) into a set of sparse vectors x_l, one for each taxonomic rank. There are L such layers. The model likelihood is a product of observation probabilities, like in the standard Naive Bayes model, but the layers are also connected by multiplication.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{l=1}^{L} \prod_{j=1}^{len(\boldsymbol{x_l})} \theta_{l,j}^{x_{l,j}} \tag{33}$$

The small difference to the Naive Bayes model in the previous section is that there are no sequence length weights and that the feature vectors are not normalized. The multiplication of layers is a

simplification because we know that taxonomic ranks are not independent. However, the model proved to be simple and effective for our purposes.

MLE for multi-layer Naive Bayes

Once the assumption of layer independence has been made, the problem simplifies to L independent Naive Bayes models with separate feature vectors and model parameters. The MLE derivation for each of these models is equivalent to the previous section. T_l is the number of features on level l.

$$\hat{\theta}_l = \frac{\sum_{i=1}^{N} t_{i,l}}{\sum_{j=1}^{T_l} \sum_{i=1}^{N} t_{i,l}} \tag{34}$$

Metagenome simulation

We chose genomes according to the CAMI2015 (www.cami-challenge.org) medium complexity toy dataset which contained 450 different strains. Because some of the strains were simulated and had no accessible genome data, we reduced the dataset to 400 genomes with corresponding accessions. These comprised both finished and draft genomes. We sampled the abundance distributions from a lognormal with expectation value one and variance one, which produced abundance value in an reasonable range and formed relative abundance by normalization (Supplementary Table 1, column S1). We derived three secondary samples (Supplementary Table 1, columns S2, S3, S4) by separately applying continuous (exponential) growth to a randomly chosen set of genomes which each constituted 100 genomes (25%) in the primary sample using the following formula.

$$abundance'(\text{genome}) = abundance(\text{gnome}) \cdot 2^{growth\_rate(\text{genome})} \tag{35}$$

We modeled the change of the community composition in reaction to variation of environmental parameters, for instance if the growth medium is altered with no space restrictions then community members will grow according to their genomic potential. In our simplified growth model we choose the growth rate uniformly at random between one and ten regardless of the actual genome. We generated three secondary abundance profiles using the described procedure. We then simulated HiSeq Illumina reads for each sample using the ART simulator with read length 150 bp, insert size 270 bp and insert size standard deviation 27 bp. This corresponds to a common experimental setting because the reads are likely to overlap in the read assembly step. We chose a large yield of 15 Gb per sample to also cover genomes with low sample abundance (see Supplementary Table 1).

Feature generation

All features are represented as separate text files, which can be compressed. Each line corresponds to a sequence but does not contain sequence identifiers. Therefore, it is required that the number and order of lines are identical in all features files.

Sequence weights

We used the following GNU awk v4.0.1 script to calculate the length of each FASTA entry which we saved as `contigs.seqlen`.

```
#!/usr/bin/awk -f
BEGIN { id="\000" } # > not allowed in FASTA header
/^>/ {
  if( id != "\000" ) {
    printf "%s\t%s\n", id, sum;
  }
  id=substr( $0, 2 );
  sum = 0;
}
! /^>/ { sum+=length($0) }
END { printf "%s\t%s\n", id, sum }
```

5-mer frequencies

We derived 5-mer frequencies for the gzip-compressed FASTA sequences using the program fasta2kmerS using the following GNU Bash syntax

```
zcat contigs.fna.gz |
fasta2kmersS -i <(cat) -f >(cat) -j 5 -k 5 -s 0 -h 0 -n 0 |
tr '\t' ' ' > contigs.kmc
```

Taxonomic annotation

We generated alignments using NCBI BLAST+/*blastn* v2.2.28+ in taxator-tk tabular format and filtered out all species level alignments using program *alignments-filter* from taxator-tk v1.3.3 which effectively removes the genomes of the same species from the reference sequences. Next we ran the program *taxator* with the LCA algorithm using only the best hits and processed the resulting GFF3 file. We used the alignment score as weight for each taxon and combined the annotations for each contig. Finally, we shortened the taxon paths using numbers and applied the described accumulation scheme to project alignment score onto higher-level taxa (see Table 1).

Average read coverage

We aligned each sample's simulated read data to the artificial contigs with Bowtie v2.2.7 and converted the resulting SAM files to sorted BAM

```
bowtie2-build contigs.fna contigs.bowtie2
bowtie2 -x contigs.bowtie2 -1 forward.fq.gz -2 reverse.fq.gz |
samtools view -@ 5 -b - < input.sam | samtools sort -@ 5 - out
```

and then calculated the average read coverage using BedTools v2.25 and GNU awk v4.0.1

```
genomeCoverageBed -ibam out.sorted.bam -g contigs.seqlen -d -split |
awk 'BEGIN{IFS=OFS=FS="\t"}
     {if($1 == last){ s+=$3; c+=1;}
     else{if(s){print last, s/c; s=$3}; c=1; last=$1}}
     END{print last, s/c}' > out.twocol.cov
```

Contigs which recruited no reads are omitted by BedTools, therefore zero values must be added afterwards by comparison to the sequence length file. Finally, we merged the coverage columns in Bash using

```
paste -d ' ' <(cut -f 2 < 1.twocol.cov) <(cut -f 2 < 2.twocol.cov) [...] > out.cov
```

Performance measures

In order to evaluate the quality of the predictions and to pick the optimal β parameter for the posterior estimation, MGLEX implements two measures: a mean squared error (MSE) and the mean pairwise coclustering (MPC) probability. Both require as input a label probability matrix which defines to which genome (column) each sequence (row) belongs, in terms of probabilities. In our simulation, the genome column corresponding to the source genome contained a one, all other columns a zero. A prediction probability matrix of the same form is required for comparison. In the case of ML predictions, this matrix also contains only ones and zeros and continuous values for the posterior estimation. Because sequences typically have different lengths, the user must provide a file with the sequence lengths (see AWK script for sequence weight file generation).

Mean squared error (MSE)

The mean squared error is the square root of the average squared difference between the label and the prediction matrix per contig (a value between zero and one). It is weighted by the length of the sequence.

$$\mathrm{MSE} = \sqrt{\frac{1}{4\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \sum_{j=1}^{M} \left(L_{i,j} - P_{i,j}\right)^2} \tag{36}$$

Here, N is the number of sequences, M the number of genomes, w is a vector with the sequence lengths, L the label probability matrix and P the prediction probability matrix.

Mean pairwise coclustering (MPC)

The mean pairwise coclustering probability reports how likely a pair of sequences chosen from any genome among the real genomes, are found in the same predicted genome. The MPC averages over both, the pairs in the genomes and the genomes, regardless of their size. Since all sequences in our evaluations have the same length, we report the unweighted version of the MPC. The MPC is a probability between zero and one. It is easier to interpret than the MSE but requires more computation because it needs to consider all possible sequence pairs.

$$\mathrm{MPC} = \frac{1}{|C|} \sum_{i=1}^{|C|} \left(\frac{1}{|C_i|(|C_i|-1)} \sum_{\substack{s_1, s_2 \in C_i \\ s_1 \neq s_2}} p(s_1|C_i) p(s_2|C_i) \right) \tag{37}$$

Here, the i^{th} genome is a set C_i which contains sequences s_i and C is a set which contains all genomes C_i.

Genome bin posterior

We calculate the bin posterior of a contig over the genome bins by normalization of the different likelihood values for each of the considered bins, so that their values sum to one. We assume, that

the bin posterior is uniform over all G genome bins, so there is no additional weighting, for instance by genome size. $\mathcal{L}(\text{genome} \mid \text{contig})$ is a vector which holds the likelihood of a specific contig for every genome bin. Then, the posterior is given by

$$P(\text{genome} \mid \text{contig}) = \frac{\mathcal{L}(\text{genome} \mid \text{contig})}{\sum_{n=1}^{G} \mathcal{L}(\text{genome}_n \mid \text{contig})} \tag{38}$$

Relative likelihood bin comparison

We derived a percentage similarity quantity S for two genome bins A and B, based on mixture likelihoods.

$$\mathrm{S}(A,B) = \sqrt[z]{\prod_{i=1}^{N} \left(\frac{2\, L_i(\theta_A)\, L_i(\theta_B)}{L_i^2(\theta_A) + L_i^2(\theta_B)} \right)^{\frac{L_i^2(\theta_A)+L_i^2(\theta_B)}{L_i(\theta_A)+L_i(\theta_b)}}} \tag{39}$$

with normalization constant

$$Z = \sum_{i=1}^{N} \frac{L_i^2(\theta_A) + L_i^2(\theta_B)}{L_i(\theta_A) + L_i(\theta_b)} \tag{40}$$

Interestingly, when we interpret this quantity as a probability, a connection to the Kullback-Leibler divergence D_{KL}, also called relative entropy, can be constructed. The Boltzmann formula (Suppl. Equation 41) establishes a general connection between entropy H and probability P.

$$\mathrm{H} = \log \mathrm{P} \tag{41}$$

When we substitute the probability P in Suppl. Equation 41 with $\mathrm{S}(A,B)$ from Suppl. Equation 39, we get

$$\begin{aligned} \mathrm{H}(A,B) &= -\frac{1}{Z} \sum_{i=1}^{N} \left(\tfrac{L_A^2+L_B^2}{L_A+L_B} \right) \log \frac{L_i^2(\theta_A) + L_i^2(\theta_B)}{2\, L_i(\theta_A)\, L_i(\theta_B)} \\ &= -\frac{1}{Z} D_{\mathrm{KL}}(\hat{L} \| L_{swap}) \end{aligned} \tag{42}$$

Suppl. Equation 42 is the negative Kullback-Leibler divergence over the sample data, which measures the loss of information when the suboptimal model with swapped parameters is used instead of the MLE parameter model, divided by the summed likelihood of the observed data.

Supplementary Tables

Supplementary Table 1: Taxa in the simulated dataset and corresponding relative abundances for the primary sample S1 and the three secondary samples S2, S3 and S4.

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Acaryochloris CCMEE 5410	0.27	0.07	0.08	0.08
Acetobacteraceae bacterium AT-5844	0.04	0.01	0.01	0.01
Acholeplasma laidlawii PG-8A	0.12	0.79	0.04	0.04
Acidaminococcus fermentans DSM 20731	0.16	0.04	0.05	0.05
Acidaminococcus BV3L6	0.29	0.08	0.21	0.92
Acidovorax ebreus TPSY	0.09	0.03	0.03	0.03
Acidovorax KKS102	0.21	0.96	1.23	0.06
Aciduliprofundum MAR08-339	1.12	0.31	0.34	0.34
Acinetobacter baumannii AB_TG2028	0.83	1.08	0.25	0.25
Acinetobacter baumannii Naval-113	0.13	0.25	0.18	0.04
Acinetobacter baumannii ZWS1122	0.05	0.06	0.01	0.01
Acinetobacter genomosp. 13TU NCTC 8102	0.06	0.02	0.02	0.12
Acinetobacter johnsonii ANC 3681	0.02	0.00	0.00	0.13
Acinetobacter nosocomialis 28F	0.07	0.02	0.02	0.02
Acinetobacter schindleri NIPH 900	0.01	0.00	0.05	0.06
Acinetobacter schindleri TG19614	0.08	0.20	0.34	0.32
Acinetobacter CIP 64.7	0.25	0.07	0.08	0.08
Actinobacillus minor NM305	0.23	0.06	0.07	0.07
Actinoplanes SE50/110	0.51	0.14	0.16	0.15
Actinopolyspora mortivallis DSM 44261	0.19	0.05	0.06	0.06
Aeromonas MDS8	0.16	0.04	0.05	0.29
Aggregatibacter actinomycetemcomitans AAS4A	0.02	0.00	0.01	0.01
Aggregatibacter actinomycetemcomitans SCC393	0.06	0.02	0.02	0.02
Alicyclobacillus acidocaldarius Tc-4-1	0.02	0.01	0.02	0.01
Alistipes CAG:53	0.14	0.04	0.28	0.04
Alloprevotella rava F0323	0.26	0.07	0.08	0.08
alpha proteobacterium LLX12A	0.07	0.02	0.02	0.02
alpha proteobacterium SCGC AAA015-O19	0.04	0.15	0.01	0.35
alpha proteobacterium SCGC AAA536-G10	0.62	0.17	0.19	5.38
Alteromonas macleodii `Ionian Sea U8'	0.05	0.04	0.01	0.01
Amphibacillus xylanus NBRC 15112	0.10	0.03	0.09	0.03
Amycolatopsis mediterranei U32	0.07	0.02	0.02	0.02
Anaerococcus hydrogenalis ACS-025-V-Sch4	0.03	0.01	0.06	0.01
Anaerococcus hydrogenalis DSM 7454	0.18	0.05	0.06	0.06
Anaplasma marginale Florida	0.01	0.00	0.01	0.00
Anaplasma marginale Gypsy Plains	0.74	0.20	0.23	4.79
Anaplasma marginale St. Maries	0.52	0.14	4.64	0.16
Anoxybacillus SK3-4	0.15	0.04	0.05	0.05
Arthrobacter FB24	0.14	0.04	0.04	0.04
Arthrobacter TB 23	0.25	0.07	0.08	0.08
Azospirillum CAG:239	0.06	0.02	0.02	0.08
Bacillus amyloliquefaciens DC-12	0.34	0.09	0.11	0.10
Bacillus anthracis A0193	0.54	3.44	1.17	3.68
Bacillus anthracis A1055	0.16	0.10	0.05	0.05
Bacillus cereus Rock1-15	0.04	0.01	0.04	0.04
Bacillus cereus Rock4-2	0.30	0.23	0.09	0.09

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Bacillus cereus VD014	0.56	0.15	0.17	0.17
Bacillus pumilus ATCC 7061	0.14	0.37	0.04	0.04
Bacillus 37MA	0.14	0.04	0.20	0.04
Bacillus EGD-AK10	0.24	0.06	0.07	0.07
Bacillus WBUNB004	0.31	0.08	0.09	0.23
Bacillus WBUNB009	0.37	0.10	0.11	0.11
Bacillus subtilis gtP20b	0.14	0.04	0.98	0.15
Bacillus subtilis S1-4	0.46	0.13	0.14	0.14
Bacillus subtilis 6051-HGW	0.10	0.03	0.03	0.03
Bacillus thuringiensis BGSC 4CC1	0.12	0.03	0.04	0.04
Bacteriovorax DB6_IX	0.11	0.03	0.03	0.03
Bacteroides faecis CAG:32	0.06	0.05	0.33	0.02
Bacteroides fragilis CAG:558	0.08	0.06	0.03	0.03
Bacteroides 4_1_36	0.20	0.05	0.29	0.06
Bacteroides CAG:443	0.27	0.07	0.08	0.08
Bacteroides CAG:714	0.04	0.01	0.01	0.03
Beijerinckia indica ATCC 9039	0.06	0.02	0.22	0.07
Bifidobacterium longum CAG:69	0.02	0.02	0.01	0.01
Bizionia argentinensis JUB59	0.31	0.09	0.10	0.27
Bordetella bronchiseptica Bbr77	0.17	0.05	0.05	0.05
Borrelia burgdorferi 29805	0.48	0.13	0.15	0.15
Brachyspira hampsonii 30599	0.10	0.03	0.03	0.03
Bradyrhizobium DFCI-1	0.11	0.06	0.03	0.03
Bradyrhizobium S23321	0.30	2.08	0.09	0.09
Bradyrhizobium WSM2793	0.03	0.06	0.01	0.01
Brevibacillus laterosporus PE36	0.05	0.14	0.02	0.39
Brevibacterium casei S18	0.40	0.54	0.12	0.12
Brevibacterium mcbrellneri ATCC 49030	0.58	3.30	0.77	0.18
Brevundimonas abyssalis TAR-001	0.37	2.08	0.11	0.11
Brevundimonas BAL3	0.18	0.05	0.06	0.06
Brucella abortus 68-3396P	0.22	0.06	0.07	0.07
Brucella abortus NI274	0.17	0.25	0.20	0.05
Burkholderia bryophila 376MFSha3.1	0.04	0.01	0.01	0.01
Burkholderia mallei 2002721280	0.25	0.07	0.08	1.08
Burkholderia pseudomallei 668	0.16	1.10	0.05	0.05
Burkholderia pseudomallei DM98	0.13	0.04	0.04	0.04
Burkholderia CCGE1001	0.09	0.03	0.90	0.03
Burkholderia WSM4176	0.05	0.01	0.02	0.02
butyrate-producing bacterium SM4/1	0.27	2.17	0.08	0.08
Butyrivibrio crossotus CAG:259	0.07	0.02	0.06	0.02
Caldicellulosiruptor bescii DSM 6725	0.51	0.14	0.16	0.16
Caldivirga maquilingensis IC-167	0.06	0.02	0.02	0.02
Candidatus Accumulibacter phosphatis UW-1	0.25	0.07	0.08	0.08
Candidatus Photodesmus katoptron Akat1	0.20	0.57	0.06	0.22
Candidatus Poribacteria WGA-A3	0.06	0.02	0.02	0.02
Candidatus Saccharibacteria RAAC3_TM7_1	0.34	0.75	0.10	0.10
Capnocytophaga F0502	0.08	0.02	0.02	0.02
Carnobacterium WN1359	0.29	0.29	0.09	0.09

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Catellicoccus marimammalium M35/04/3	0.33	0.57	0.10	0.36
Chitinophaga pinensis DSM 2588	0.24	0.06	0.33	0.07
Chlamydia psittaci WC	0.05	0.01	0.02	0.20
Chlamydia trachomatis IU888	0.02	0.00	0.02	0.01
Chlamydia trachomatis L2b/Ams2	0.05	0.01	0.01	0.04
Chlamydia trachomatis RC-J/953	0.72	0.20	0.22	0.22
Chloroflexi bacterium oral isolate Chl1-2	0.35	0.09	0.11	0.11
Chloroflexi bacterium SCGC AB-629-P13	0.32	0.09	0.10	1.65
Citrobacter rodentium ICC168	0.08	0.02	0.02	0.02
Citrobacter KTE151	0.04	0.04	0.01	0.25
Clostridium acetobutylicum EA 2018	0.18	0.05	0.06	1.13
Clostridium carboxidivorans P7	0.23	0.25	2.18	0.07
Clostridium ATCC BAA-442	0.32	0.09	0.10	1.42
Clostridium CAG:269	0.38	0.10	0.83	1.36
Clostridium CAG:452	0.21	0.06	0.06	0.06
Clostridium CAG:567	0.44	0.12	0.14	0.53
Clostridium SY8519	0.10	0.03	0.03	0.03
Clostridium tyrobutyricum DSM 2637/ATCC 25755/JCM 11008	0.88	0.24	4.86	0.27
Collimonas fungivorans Ter331	0.19	0.05	0.33	0.06
Coprococcus comes CAG:19	0.10	0.03	0.09	0.03
Corynebacterium pseudotuberculosis 316	0.02	0.00	0.00	0.00
Corynebacterium pseudotuberculosis Cp162	0.08	0.02	0.02	0.20
Corynebacterium pseudotuberculosis I19	0.07	0.02	0.02	0.02
Corynebacterium KPL1855	0.82	4.06	0.25	0.60
Corynebacterium KPL1859	0.09	0.09	0.23	0.03
Corynebacterium KPL1998	0.09	0.03	0.03	0.26
Cronobacter sakazakii 701	0.11	0.03	0.03	0.03
Cupriavidus basilensis B-8	0.11	0.10	0.03	0.03
Cyanothece CCY0110	0.08	0.02	0.03	0.02
Cyclobacterium qasimii M12-11B	0.13	0.04	0.04	0.50
Desulfococcus oleovorans Hxd3	0.10	0.03	0.12	0.03
Desulfovibrio aespoeensis Aspo-2	0.22	0.06	0.07	0.07
Desulfurivibrio alkaliphilus AHT2	0.18	0.05	0.05	0.05
Dictyoglomus turgidum DSM 6724	0.39	0.11	0.12	0.12
Eggerthia catenaformis OT 569/DSM 20559	0.33	0.09	0.10	0.40
Emticicia oligotrophica DSM 17448	0.31	0.09	0.10	0.10
Enterobacter R4-368	0.07	0.02	0.02	0.02
Enterococcus flavescens ATCC 49996	0.08	0.02	0.02	0.02
Enterococcus GMD4E	1.14	0.31	0.35	0.35
Enterovibrio norvegicus FF-162	0.49	0.13	0.15	0.15
Erysipelotrichaceae bacterium 5_2_54FAA	0.27	0.15	0.08	0.08
Erythrobacter litoralis HTCC2594	0.86	0.23	0.26	0.26
Exiguobacterium pavilionensis RW-2	0.11	0.36	0.03	0.03
Facklamia ignava CCUG 37419	0.58	0.51	0.39	0.18
Faecalibacterium prausnitzii A2-165	0.08	0.02	0.02	0.02
Finegoldia magna BVS033A4	0.07	0.02	0.07	0.02
Firmicutes bacterium ASF500	0.09	0.02	0.03	0.03
Firmicutes bacterium CAG:170	0.17	0.05	0.05	0.05

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Fischerella thermalis PCC 7521	0.13	0.04	0.04	0.14
Flavobacteriaceae bacterium S85	0.41	0.11	0.13	0.13
Flavobacterium B17	0.14	0.04	0.04	0.04
Formosa AK20	0.64	0.18	0.20	0.20
Francisella tularensis 80700075	0.08	0.07	0.03	0.12
Frankia alni ACN14a	0.33	0.09	0.10	0.73
gamma proteobacterium IMCC2047	0.09	0.02	0.03	0.03
Gardnerella vaginalis 0288E	0.04	0.01	0.01	0.01
Gardnerella vaginalis 1500E	0.27	0.07	0.08	0.08
Geobacillus JF8	0.11	0.72	0.04	0.03
Gillisia marina	0.41	0.11	0.13	0.13
Glaciecola polaris LMG 21857	0.26	0.07	0.08	0.08
Glaciecola 4H-3-7+YE-5	0.33	0.21	0.10	0.10
Gordonia effusa NBRC 100432	0.09	0.03	0.45	0.03
Gordonia sihwensis NBRC 108236	0.12	0.24	0.04	0.11
Haemophilus aegyptius ATCC 11116	0.48	0.13	0.15	0.15
Haemophilus somnus 129PT	0.31	0.09	0.10	0.10
Haemophilus sputorum HK 2154	0.94	0.26	0.29	0.29
Haloferax BAB2207	0.43	0.12	0.13	0.13
Halomonas KM-1	0.13	0.03	0.04	1.02
Halorhabdus utahensis DSM 12940	0.01	0.00	0.04	0.00
Haloterrigena limicola JCM 13563	0.04	0.01	0.01	0.01
Helicobacter hepaticus ATCC 51449	0.38	0.10	3.30	0.39
Herbaspirillum B39	0.41	0.11	0.12	0.12
Ignavibacterium album JCM 16511	0.39	0.11	0.12	0.12
Isoptericola variabilis 225	0.20	0.05	0.06	0.06
Janibacter HTCC2649	0.43	0.12	0.13	0.95
Kingella kingae PYKK081	0.19	0.05	0.12	0.06
Klebsiella pneumoniae UHKPC01	0.18	1.40	0.06	0.06
Klebsiella pneumoniae UHKPC02	0.14	0.04	1.04	0.04
Klebsiella pneumoniae UHKPC40	0.19	0.05	0.06	1.46
Ktedonobacter racemifer DSM 44963	0.14	0.04	0.04	0.04
Laceyella sacchari 1-1	0.08	0.02	0.02	0.21
Lachnospiraceae bacterium 2_1_46FAA	0.54	1.60	0.17	0.16
Lachnospiraceae bacterium 3-2	0.33	0.09	0.79	0.63
Lachnospiraceae bacterium 5_1_57FAA	0.04	0.22	0.01	0.37
Lachnospiraceae oral taxon 107 str. F0167	0.19	0.35	0.06	0.06
Lactobacillus acidipiscis KCTC 13900	0.04	0.01	0.01	0.01
Lactobacillus acidophilus 30SC	0.38	0.10	0.77	0.11
Lactobacillus acidophilus ATCC 4796	0.04	0.01	0.01	0.01
Lactobacillus casei 21/1	0.22	0.43	0.07	0.07
Lactobacillus casei Lpc-37	0.21	0.06	0.06	0.06
Lactobacillus delbrueckii ATCC BAA-365	0.03	0.01	0.01	0.01
Lactobacillus delbrueckii DSM 20072	0.32	0.09	1.90	2.76
Lactobacillus fermentum CECT 5716	0.42	0.11	0.13	0.13
Lactobacillus helveticus CNRZ32	0.08	0.20	0.02	0.02
Lactobacillus helveticus R0052	0.10	0.03	0.53	0.03
Lactobacillus iners ATCC 55195	0.06	0.02	0.02	0.06

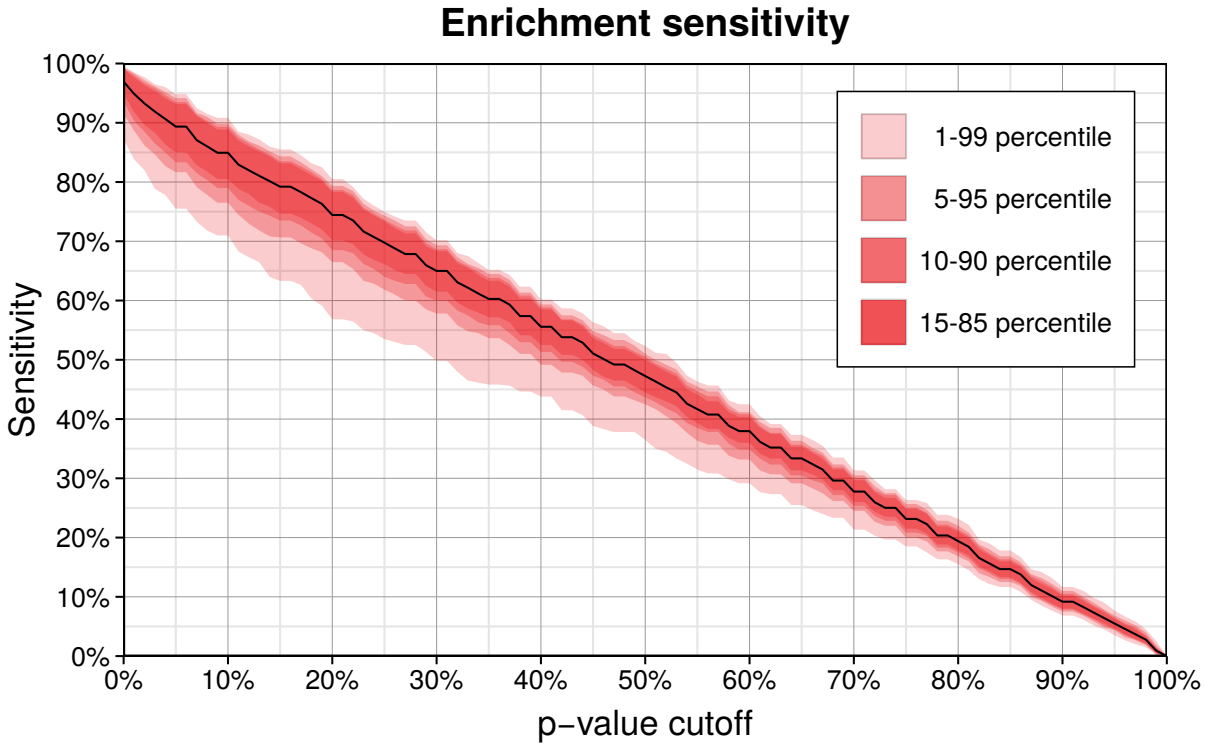
Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Lactobacillus iners LactinV 01V1-a	0.11	0.03	0.03	0.03
Lactobacillus plantarum 2165	0.85	0.23	0.26	0.54
Lactobacillus reuteri CF48-3A	0.66	0.18	0.20	0.20
Lactobacillus reuteri MM4-1A	0.23	1.14	0.07	0.07
Lactobacillus salivarius GJ-24	0.42	0.12	0.13	1.37
Lactobacillus ASF360	0.28	0.08	1.93	0.09
Legionella pneumophila str. 121004	0.05	0.06	0.01	0.29
Leifsonia xyli subxyli str. CTCB07	0.04	0.01	0.01	0.01
Leptospira borgpetersenii 200801910	0.20	0.05	0.06	0.06
Leptospira borgpetersenii 200901122	0.24	0.17	0.38	0.16
Leptospira interrogans Fiocruz R154	0.15	0.04	0.05	0.05
Leptospira interrogans L1207	0.11	0.03	0.04	0.03
Leptospira santarosai Oregon	0.59	3.58	0.18	0.18
Leptospira santarosai 2000027870	0.12	0.03	0.24	0.04
Leptospira santarosai HAI1380	0.13	0.04	0.04	0.04
Leuconostoc argentinum KCTC 3773	0.13	0.04	0.09	0.53
Leuconostoc citreum LBAE C10	0.02	0.01	0.20	0.07
Loktanella cinnabarina LL-001	0.23	0.06	0.07	0.07
Loktanella hongkongensis DSM 17492	0.18	0.05	0.06	0.53
Mannheimia haemolytica USDA-ARS-USMARC-183	0.19	0.05	0.06	0.06
marine gamma proteobacterium HTCC2080	0.10	0.06	0.30	0.03
Marinimicrobia bacterium SCGC AAA298-D23	0.26	0.65	0.08	0.08
Marinimicrobia bacterium SCGC AB-629-J13	0.26	0.07	0.08	0.08
Marinobacter EVN1	0.05	0.01	0.43	0.15
Megasphaera genomosp. type_1 str. 28L	0.14	0.04	0.47	0.04
Melissococcus plutonius DAT561	0.39	0.58	0.12	0.12
Mesoflavibacter zeaxanthinifaciens S86	0.11	0.03	0.21	0.03
Mesorhizobium LNHC229A00	0.16	1.18	0.11	0.05
Mesorhizobium LSHC416B00	0.04	0.01	0.01	0.01
Mesorhizobium LSJC264A00	0.04	0.08	0.01	0.01
Methanobrevibacter smithii TS146D	0.14	0.04	0.80	0.04
Methanobrevibacter smithii TS147C	0.14	0.04	0.04	0.11
Methanobrevibacter smithii TS95A	0.09	0.02	0.17	0.03
Methanocella arvoryzae MRE50	0.05	0.13	0.02	0.13
Methanosphaera stadtmanae DSM 3091	0.18	0.05	0.06	0.05
Methylobacterium extorquens PA1	0.10	0.78	0.03	0.03
Methyloglobulus morosus KoM1	0.19	0.05	0.06	1.18
Methylotenera versatilis 301	0.06	0.02	0.02	0.05
Methyloversatilis universalis EHg5	0.17	0.05	0.05	0.05
Microbacterium barkeri 2011-R4	0.12	0.03	0.04	0.04
Microbacterium 11MF	0.13	0.08	0.37	0.04
Microbacterium TS-1	0.10	0.03	0.03	0.03
Mobiluncus curtisii ATCC 43063	0.39	0.11	0.12	0.12
Mycobacterium abscessus 3A-0930-R	0.03	0.01	0.01	0.01
Mycobacterium abscessus 5S-0422	0.58	0.16	0.37	0.99
Mycobacterium abscessus M139	0.26	0.07	0.08	0.08
Mycobacterium chubuense NBB4	0.18	0.05	0.06	0.05
Mycobacterium intracellulare MOTT-02	0.19	0.05	0.06	0.12

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Mycoplasma gallisepticum NC08_2008.031-4-3P	0.05	0.02	0.02	0.02
Mycoplasma gallisepticum NY01_2001.047-5-1P	0.27	0.07	0.08	0.08
Neisseria gonorrhoeae PID18	0.13	0.03	0.04	0.04
Neisseria gonorrhoeae SK-92-679	0.13	0.04	0.04	0.04
Neisseria meningitidis NM1476	0.18	0.16	0.05	0.05
Neisseria meningitidis NM3223	0.16	0.05	0.05	0.15
Neisseria meningitidis NM604	0.27	0.07	0.08	0.08
Neisseria sicca 4320	0.09	0.23	0.03	0.03
Niabella aurantiaca DSM 17617	0.15	0.23	0.32	0.04
Nitrolancea hollandica Lb	0.36	0.78	1.24	0.11
Nocardia tenerifensis NBRC 101015	0.40	0.11	0.12	1.67
Nocardiopsis CNS639	0.74	0.20	0.23	0.23
Nonomuraea coxensis DSM 45129	0.69	1.26	0.21	1.07
Oceanicaulis HTCC2633	0.19	0.05	0.34	0.06
Oceanobacillus kimchii X50	0.74	2.17	0.23	0.86
Octadecabacter arcticus 238	0.06	0.02	0.02	0.02
Paenibacillus alvei TS-15	0.19	0.05	1.36	0.06
Paenibacillus larvae BRL-230010	0.03	0.01	0.01	0.01
Paenibacillus Aloe-11	0.04	0.01	0.01	0.01
Pantoea AS-PWVM4	0.09	0.02	0.03	0.03
Parabacteroides ASF519	0.19	0.05	0.06	0.85
Parascardovia denticolens IPLA 20019	0.42	0.11	2.48	3.38
Parasutterella excrementihominis CAG:233	0.72	0.20	0.22	1.60
Patulibacter americanus DSM 16676	0.07	0.02	0.02	0.55
Patulibacter medicamentivorans	0.45	0.12	1.09	0.14
Pediococcus acidilactici D3	0.07	0.05	0.02	0.02
Pelosinus fermentans A11	0.04	0.01	0.23	0.03
Peptoclostridium difficile P20	0.12	0.03	0.04	0.53
Peptoclostridium difficile P48	0.04	0.01	0.01	0.08
Peptoclostridium difficile P53	0.24	0.07	1.05	0.07
Polynucleobacter necessarius QLW-P1DMWA-1	0.09	0.02	0.03	0.03
Porphyromonas gingivalis JCVI SC001	0.17	0.05	0.23	1.25
Porphyromonas gingivalis W50	0.81	0.22	0.25	0.25
Porphyromonas macacae DSM 20710/JCM 13914	0.11	0.03	0.03	0.03
Prevotella salivae DSM 15606	0.04	0.01	0.01	0.01
Prevotella C561	0.03	0.19	0.01	0.01
Prevotella CAG:1185	0.39	0.11	3.30	0.12
Prevotella CAG:592	0.35	1.04	1.14	0.11
Prevotella CAG:617	0.32	0.09	0.10	0.91
Prevotella CAG:755	0.14	0.04	0.04	0.04
Prevotella CAG:873	0.07	0.02	0.02	0.02
Pseudomonas aeruginosa BWHPSA006	0.10	0.03	0.03	0.03
Pseudomonas aeruginosa LESB58	0.23	0.20	1.02	0.07
Pseudomonas aeruginosa PABL056	0.09	0.03	0.03	0.03
Pseudomonas mendocina ymp	0.10	0.03	0.03	0.18
Pseudomonas CF161	0.07	0.02	0.02	0.02
Pseudomonas EGD-AK9	0.04	0.01	0.03	0.01
Pseudomonas M47T1	0.28	0.08	0.09	0.39

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Pseudomonas TJI-51	0.03	0.01	0.01	0.01
Pseudomonas syringae pv. lachrymans M302278	0.25	0.70	1.23	0.08
Psychrobacter PRwf-1	0.03	0.01	0.01	0.01
Pyrobaculum aerophilum str. IM2	0.32	0.09	0.10	0.10
Pyrobaculum calidifontis JCM 11548	0.03	0.01	0.01	0.07
Pyrococcus furiosus COM1	0.36	0.10	0.11	0.11
Ralstonia solanacearum Po82	0.18	0.05	0.06	0.05
Renibacterium salmoninarum ATCC 33209	0.49	0.13	0.15	0.15
Rhizobium etli Brasil 5	0.02	0.01	0.01	0.01
Rhizobium phaseoli Ch24-10	0.08	0.33	0.02	0.02
Rhizobium IRBG74	0.07	0.12	0.02	0.70
Rhodobacter SW2	0.17	0.05	0.05	0.05
Rhodobacter sphaeroides ATCC 17029	0.47	0.13	0.14	2.85
Rhodobacteraceae bacterium KLH11	1.39	3.41	3.28	0.42
Rhodococcus rhodnii LMG 5362	0.24	1.34	0.07	0.22
Rhodococcus 29MFTsu3.1	0.06	0.02	0.02	0.02
Rhodococcus P27	0.22	0.06	0.07	0.07
Rhodopirellula baltica SWK14	0.55	0.15	0.17	0.17
Rhodopseudomonas palustris BisB5	0.28	0.08	0.08	0.08
Rhodospirillum rubrum ATCC 11170	0.02	0.01	0.01	0.05
Rickettsia helvetica C9P9	0.19	0.05	0.06	0.06
Rickettsia rickettsii str. `Sheila Smith'	0.49	0.32	1.97	1.27
Riemerella anatipestifer RA-YM	0.08	0.02	0.03	0.65
Rudanella lutea DSM 19387	0.72	0.20	0.22	0.22
Ruminiclostridium thermocellum ATCC 27405	0.42	0.11	0.13	0.13
Ruminiclostridium thermocellum YS	0.55	0.15	1.98	0.17
Ruminococcus CAG:382	0.10	0.03	0.03	0.03
Ruminococcus CAG:579	0.88	2.32	0.27	0.27
Saccharomonospora cyanea NA-134	0.13	0.04	0.04	0.04
Salinispora arenicola CNT849	0.86	0.23	0.26	0.26
Salinispora arenicola CNY234	0.61	0.17	0.19	0.19
Salinispora pacifica CNY330	0.52	0.72	0.16	0.16
Salmonella enterica SA-2	0.05	0.01	0.02	0.02
Salmonella enterica CFSAN001588	0.13	0.04	0.89	0.04
Selenomonas noxia ATCC 43541	0.06	0.02	0.02	0.02
Shewanella frigidimarina NCIMB 400	0.04	0.04	0.04	0.01
Shigella boydii 965-58	0.18	0.05	0.58	0.05
Shigella dysenteriae CDC 74-1112	0.20	1.73	0.50	0.45
Shigella flexneri 1485-80	0.11	0.03	0.03	0.03
Shigella flexneri 2930-71	0.11	0.03	0.03	0.03
Simonsiella muelleri ATCC 29453	0.31	0.08	0.09	0.09
Sphingomonas melonis DAPP-PG 224	0.48	0.13	0.15	0.15
Sphingopyxis MC1	0.07	0.07	0.02	0.08
Staphylococcus hominis SK119	0.09	0.03	0.03	0.08
Streptococcus agalactiae GB00264	0.09	0.05	0.03	0.03
Streptococcus agalactiae MRI Z1-022	0.14	0.04	0.04	0.04
Streptococcus agalactiae MRI Z1-202	0.38	0.90	0.12	0.12
Streptococcus anginosus F0211	0.10	0.03	0.37	0.03

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Streptococcus equi	0.37	0.10	1.54	0.11
Streptococcus equi SzS31A1	0.31	0.08	0.09	0.27
Streptococcus ferus DSM 20646	0.15	0.04	0.05	0.05
Streptococcus gordonii CH1	0.26	0.07	1.11	0.08
Streptococcus iniae 9117	0.01	0.00	0.01	0.09
Streptococcus intermedius ATCC 27335	3.19	0.87	0.98	0.97
Streptococcus mutans KK23	0.62	0.17	0.19	0.19
Streptococcus mutans SM6	0.10	0.03	0.03	0.03
Streptococcus pseudoporcinus LQ 940-04	0.03	0.01	0.01	0.01
Streptococcus salivarius 57.I	0.17	0.14	0.05	0.05
Streptococcus sanguinis SK340	0.02	0.06	0.01	0.03
Streptococcus sobrinus DSM 20742/ATCC 33478	0.42	0.11	0.13	0.29
Streptococcus sobrinus TCI-367	1.65	0.45	0.50	0.50
Streptococcus sobrinus TCI-98	0.23	0.06	0.34	0.07
Streptococcus I-P16	0.05	0.01	0.01	0.01
Streptococcus SK140	0.43	0.12	0.13	0.13
Streptococcus suis YB51	0.42	0.11	0.13	0.13
Streptomyces acidiscabies 84-104	0.22	0.26	0.07	0.07
Streptomyces albulus CCRC 11814	0.40	0.11	0.92	0.12
Streptomyces pristinaespiralis ATCC 25486	0.11	0.27	0.03	0.09
Streptomyces CNQ766	0.17	0.05	0.15	0.05
Streptomyces sulphureus DSM 40104	0.13	0.11	0.45	0.04
Streptomyces violaceusniger Tu 4113	0.27	0.08	0.08	0.08
Succinatimonas hippei YIT 12066	0.10	0.03	0.03	0.08
Sulfolobus islandicus REY15A	0.10	0.03	0.44	0.03
Synechococcus PCC 7336	0.42	0.12	0.13	0.13
Synechocystis PCC 6803	0.04	0.01	0.30	0.01
Synechocystis PCC 7509	0.04	0.05	0.01	0.01
Thauera linaloolentis 47Lol/DSM 12138	0.28	0.08	0.08	0.08
Thermococcus onnurineus NA1	0.15	0.18	0.09	0.04
Thermoplasmatales archaeon I-plasma	0.13	0.04	0.04	0.04
Thermosphaera aggregans DSM 11486	0.69	0.74	0.21	1.64
Thermotoga elfii NBRC 107921	0.16	0.04	0.24	0.91
Thermotoga EMP	0.51	0.14	0.16	3.02
Thermus CCB_US3_UF1	0.17	0.05	0.05	0.05
Thioalkalivibrio AKL6	0.36	0.10	0.11	0.11
Thioalkalivibrio ALE20	0.38	0.61	0.12	0.11
Thioalkalivibrio ALJ10	0.60	2.66	0.19	0.18
Thioalkalivibrio ALJ12	0.81	0.22	0.65	0.25
Thioalkalivibrio ALJ24	0.48	3.07	0.15	0.15
Thioalkalivibrio ALJ5	0.10	0.03	0.28	0.47
Thioalkalivibrio ALJ9	0.10	0.03	0.03	0.03
Tyzzerella nexilis DSM 1787	0.16	0.04	0.05	0.05
uncultured archaeon A07HR60	0.67	0.18	2.13	5.67
Ureaplasma urealyticum ATCC 27814	0.32	0.09	1.24	0.10
Variovorax paradoxus S110	0.08	0.02	0.11	0.07
Verrucomicrobium 3C	1.48	7.16	2.75	0.45
Vibrio cholerae HC-50A2	0.16	0.04	0.05	0.05

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Vibrio cholerae HE39	0.08	0.02	0.02	0.02
Vibrio cholerae O1 str. 2009V-1085	0.03	0.01	0.08	0.01
Vibrio crassostreae 9ZC88	0.18	0.05	0.05	0.05
Vibrio gazogenes ATCC 43941	0.96	2.34	0.29	0.29
Vibrio nigripulchritudo ENn2	0.71	0.20	0.22	0.22
Vibrio nigripulchritudo SFn135	0.22	0.06	1.80	0.07
Vibrio nigripulchritudo SOn1	0.46	0.13	0.14	0.14
Weissella koreensis KACC 15510	0.25	0.38	0.25	0.08
Wolbachia endosymbiont JHB	0.39	0.11	0.12	0.12
Xanthomonas axonopodis IBSBF 614	0.06	0.02	0.10	0.02
Xanthomonas axonopodis UA306	0.15	0.04	0.05	0.05
Xanthomonas campestris NCPPB 2005	0.17	0.05	0.05	0.05
Xanthomonas oryzae BLS256	0.18	0.05	0.06	0.06
Xanthomonas SHU166	0.13	0.04	0.04	0.04
Xylella fastidiosa 32	0.18	0.23	0.05	0.05
Yersinia frederiksenii ATCC 33641	0.22	1.08	0.07	1.20
Yersinia pseudotuberculosis B-6863	0.22	0.06	0.07	0.07
Yersinia pseudotuberculosis B-6864	0.12	1.02	0.13	0.19



Supplementary Figure 1: Genome enrichment for 400 genomes in the three-fold cross-validation. For each genome, we measured the sensitivity, the percentage of each genome in the enriched sample, after filtering by a p-value cutoff and summing over the three data partitions. The solid lines shows the resulting average sensitivity over all 400 genomes. The variability between genomes is shown as quantiles in red.