Supplementary Materials

1 Flow chart of RAIDA

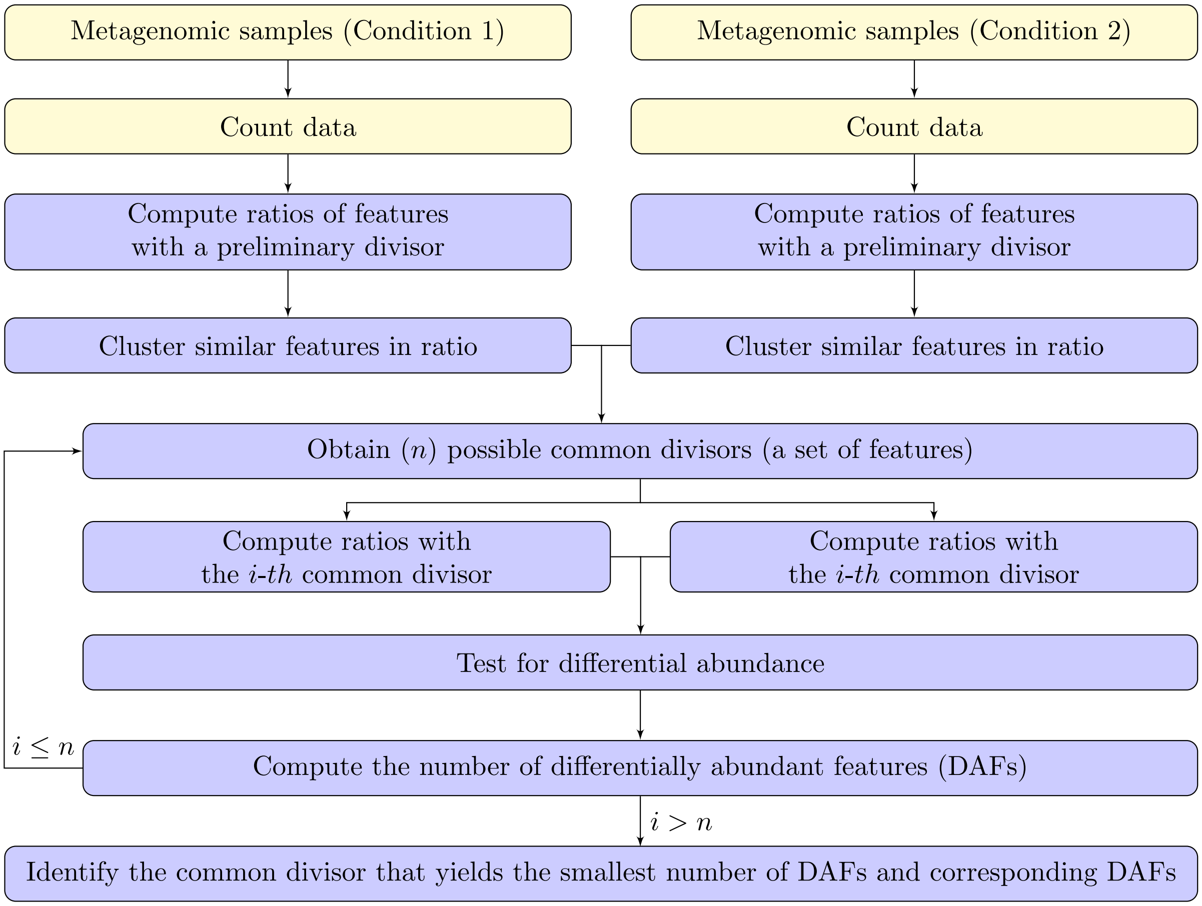
Metagenomic samples (Condition 1)
Metagenomic samples (Condition 2)
Count data
Count data
Compute ratios of features with a preliminary divisor
Compute ratios of features with a preliminary divisor
Cluster similar features in ratio
Cluster similar features in ratio
Obtain (n) possible common divisors (a set of features)
Compute ratios with the *i-th* common divisor
Compute ratios with the *i-th* common divisor
Test for differential abundance
$i \leq n$
Compute the number of differentially abundant features (DAFs)
$i > n$
Identify the common divisor that yields the smallest number of DAFs and corresponding DAFs

Figure S1: Flow chart of RAIDA's workflow. The light blue colored blocks are implemented by RAIDA.

2 Proposition:

Proposition: Under the assumption that most features are not differentially abundant, the minimum number of DAFs is achieved when a non-DAF or a group of non-DAFs is used as a common divisor.

Proof: Let g_{ci} be the i^{th} group of features similar in terms of abundance in condition c, n_i the number of features in the i^{th} group, and k_i the equivalent factor for the i^{th} group in different conditions (e.g. for two conditions $g_{1i} = k_i g_{2i}$). Note that $k_i = 1$ for non-differentially abundant groups. Assume there exists m groups in which the first ℓ groups are differentially abundant and let $n = \sum_i^m n_i$. Then, the number of DAFs is $n - n_i$ if $g_{ci} \in \{g_{c1}, \ldots, g_{c\ell}\}$ is used as a common divisor in condition c, assuming $k_i \neq k_j$ for $i \neq j$ and $i, j \leq \ell$. On the other hand, the number of DAFs is $n - \sum_{i=\ell+1}^{m} n_i$ if $g_{ci} \in \{g_{c\ell+1}, \ldots, g_{cm}\}$ is used as a common divisor since $k_i = 1$ for $i \in \{\ell + 1, \ldots, m\}$. Because $\sum_{i=1}^{\ell} n_i < \sum_{i=\ell+1}^{m} n_i$ by the assumption, $(n - n_i) > n - \sum_{i=\ell+1}^{m} n_i$. That is, the minimum number of DAFs is achieved when a group of features that belongs to non-DAFs is used as a common divisor. Note that if $n_i \in \{n_1, \dots, n_\ell\}$, the smallest number of DAFs achievable is $n - \sum_{i=1}^{\ell} n_i$, which occurs only if $k_i = k_j$ for $i, j \leq \ell$, and is still greater than $n - \sum_{i=\ell+1}^{m} n_i$ by the assumption. □

3 Simulation settings

In simulation studies, we generated counts using a zero-inflated negative binomial given by

$$c_{ij} = \xi_i \, \mathrm{I}(c_{ij} = 0) + (1 - \xi_i)\mathrm{NB}(w_i u_i s_j, \gamma_i) \, \mathrm{I}(c_{ij} \geq 0).$$

where ξ_i is the proportion of feature i in the false zero state, w_i the difference factor of feature i between two conditions, u_i the expected common mean count of feature i across all samples in the two conditions, s_j the scaling factor of sample j, I the indicator function and $\mathrm{NB}(\mu, \gamma)$ the probability mass function of a negative binomial with the parameters of mean μ and size γ, which is given by

$$P_{NB}(C = c) = \frac{\Gamma(c + \gamma)}{\Gamma(\gamma)\Gamma(c + 1)} \left(\frac{\gamma}{\gamma + \mu}\right)^{\gamma} \left(\frac{\mu}{\gamma + \mu}\right)^{c}.$$

In order to set the range of the size parameter for the simulation studies, we obtained the frequencies of the size from non-zero counts of three real datasets: bulk of soil, human saliva and Sargasso sea. These data are publicly available in MG-RAST (http://metagenomics.anl.gov/). The identification numbers for samples in each dataset are provided in Table S1. The values of the size are highly frequent in the range between 0.5 and 10 as shown in Figure S2. The frequencies of the size are also high at around 100 and between 330 and 360. However, at high values of the size the standard deviation is smaller than the mean as long as the mean is greater than 1 and the probability of having zero is very small unless the mean is very small. That is, it is relatively ease to distinguish two NB distributions with high values of the size. Therefore, we randomly selected the values of the size just from (0.1, 10). Here we assumed equal probability (or uniform distribution) in this interval even though the frequencies of the size are right skewed. The reasons are, to some extent, to avoid overdrawing the low values (especially $1 \leq \gamma \leq 3$) of the size and account for the over-estimated values of the size due to the use of only nonzero counts in the estimation of the size. That is, we assign probabilities of high values of the size to those of the size between 3 and 10 to avoid overdrawing of the values between 1 and 3. We increase probabilities of very low values of the sizes because some zeros are parts of a NB distribution so we would have estimated smaller values of the size if we could correctly include the zeros in the estimation of the size.

Table S1: Information about the real data used to estimate the size parameter.

	No. of samples	MG-RAST ID
Bulk soil	13	4449249.3, 4449252.3, 4449255.3, 4449256.3, 4449356.3, 4449357.3, 4449359.3, 4449360.3, 4449362.3, 4449363.3, 4449364.3, 4449365.3, 4449877.3
Human saliva	12	4472804.3, 4472821.3, 4473347.3, 4473348.3, 4473365.3, 4473372.3, 4473378.3, 4473389.3, 4473411.3, 4473417.3, 4473438.3, 4478542.3
Sargasso sea	21	4449104.3, 4494598.3, 4494599.3, 4494600.3, 4494602.3, 4494603.3, 4494604.3, 4494605.3, 4494606.3, 4494607.3, 4494608.3, 4494609.3, 4539503.3, 4539504.3, 4539506.3, 4539507.3, 4539508.3, 4539509.3, 4539511.3, 4539512.3, 4539513.3

For the range of the mean parameter, we used the products of randomly selected values of the three parameters: the expected common mean count u, the difference factor w and the scaling factor s. The range of u was set to any integer between 1 and 5, that of w any integer between 2 and 6, and that of s any integer

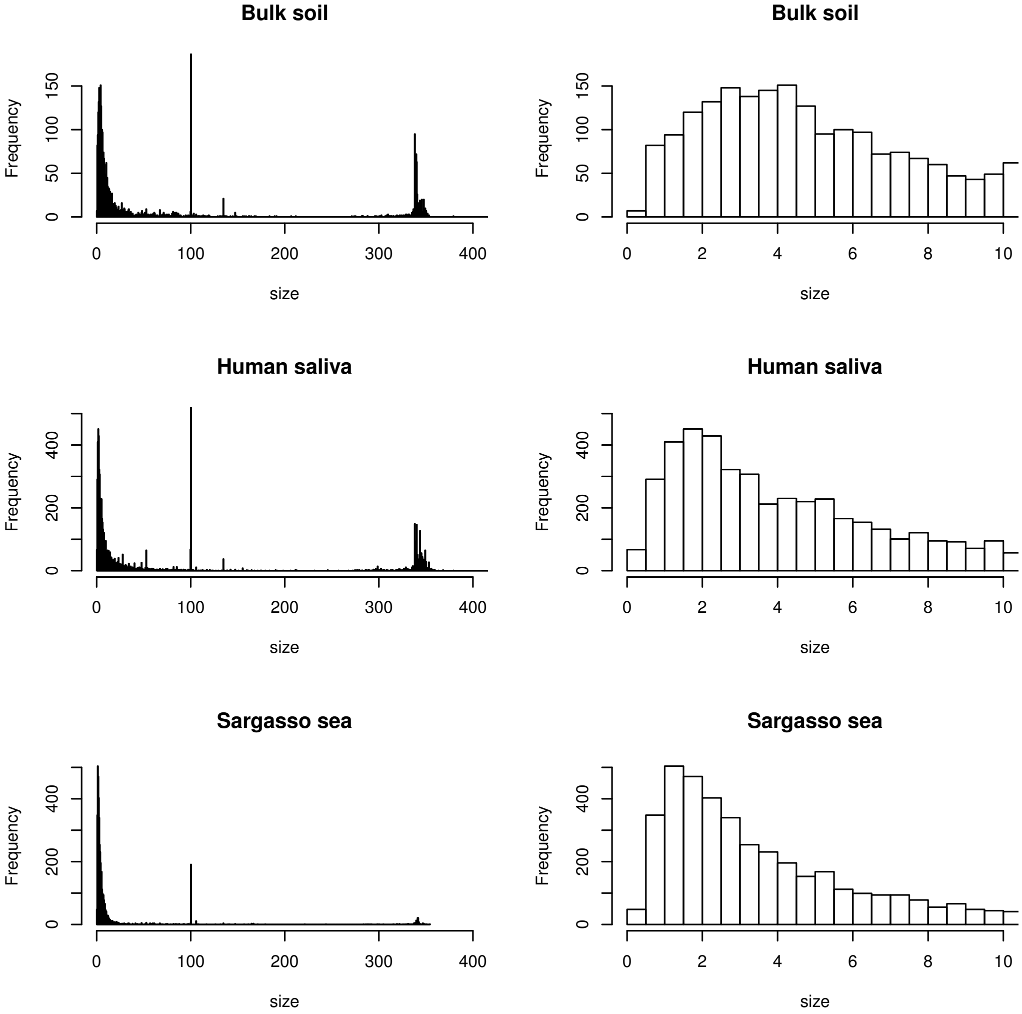

Figure S2: Frequencies of the size obtained from three real datasets: bulk of soil, human saliva and Sargasso sea. The right plots are magnified versions of the left plots, focusing on the frequency of the size between 0 and 10.

between 1 and 10. In pilot studies, we used wider ranges of the parameters in order to reproduce counts in the real datasets. However, the effects of wider ranges of u and s on identifying DAFs were very small. Most of the time all the methods used in this paper identified DAFs correctly when u is greater than or equal to 7 folds. Thus, we reduced the ranges to show the difference in performance clearer in the comparison of the methods. Note w is a difference factor between two conditions, so $w = 1$ for the non-DAFs. For the proportion of feature i in the false zero state, we used $\xi_i = \exp(-1/\mu_i) + \text{int}(0, 2)/10$, where $\text{int}(a, b)$ denote any randomly selected integer between a and b. Summary of the ranges of all the parameters is given in Table S2.

Table S2: The ranges of parameters used in the simulation study. The subscript i represents the ith feature, and the subscript j represents the jth sample. $\text{int}(a, b)$ denotes any randomly selected integer between a and b, and $\text{unif}(a, b)$ denotes any randomly selected real between a and b.

ξ_i	u_i	γ_i	w_i	s_j
$\exp(-1/\mu_i) + \text{int}(0, 2)/10$	$\text{int}(1, 5)$	$\text{unif}(0.1, 10)$	$\text{int}(2, 6)$	$\text{int}(1, 10)$

3.1 Setting for the balanced and unbalanced conditions

Balanced conditions: 1^{st} half of DAFs with $w_i = \text{int}(2, 6)$ and the other half with $w_i = 1$ in one condition, and 1^{st} half with $w_i = 1$ and the other half with $w_i = \text{int}(2, 6)$ in the other condition.

Unbalanced conditions: DAFs with $w_i = \text{int}(2, 6)$ in one condition, and DAFs with $w_i = 1$ in the other condition.

4 Results for the simulation study

4.1 10% balanced and unbalanced DAFs

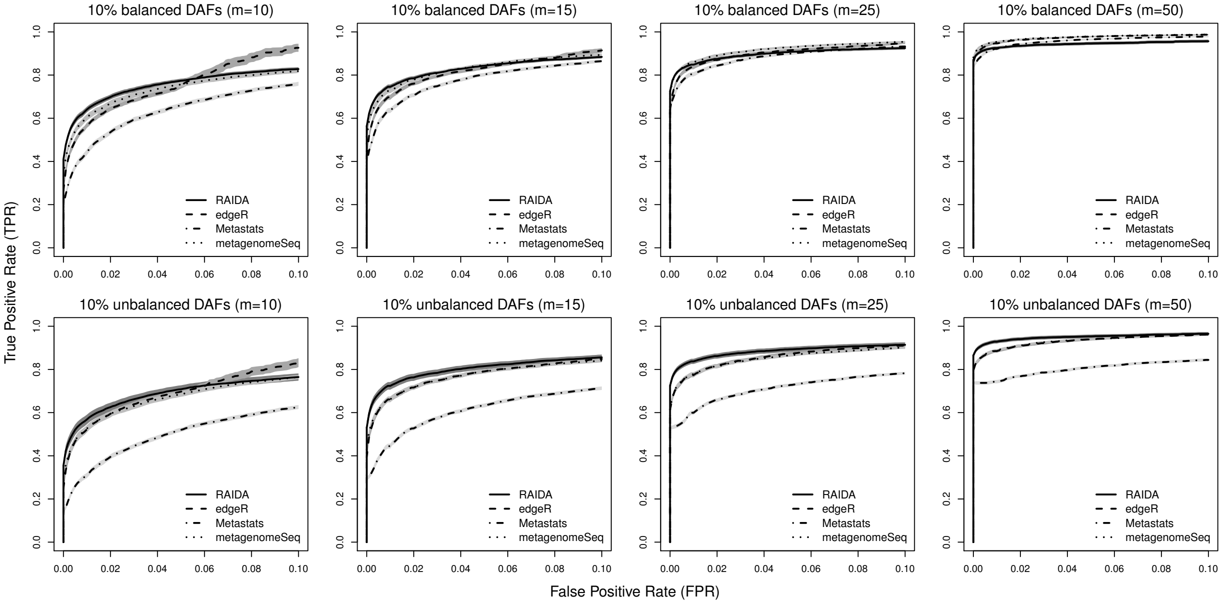

Figure S3: Partial of mean ROC curves for 10% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The shades around the lines are 95% confidence bands.

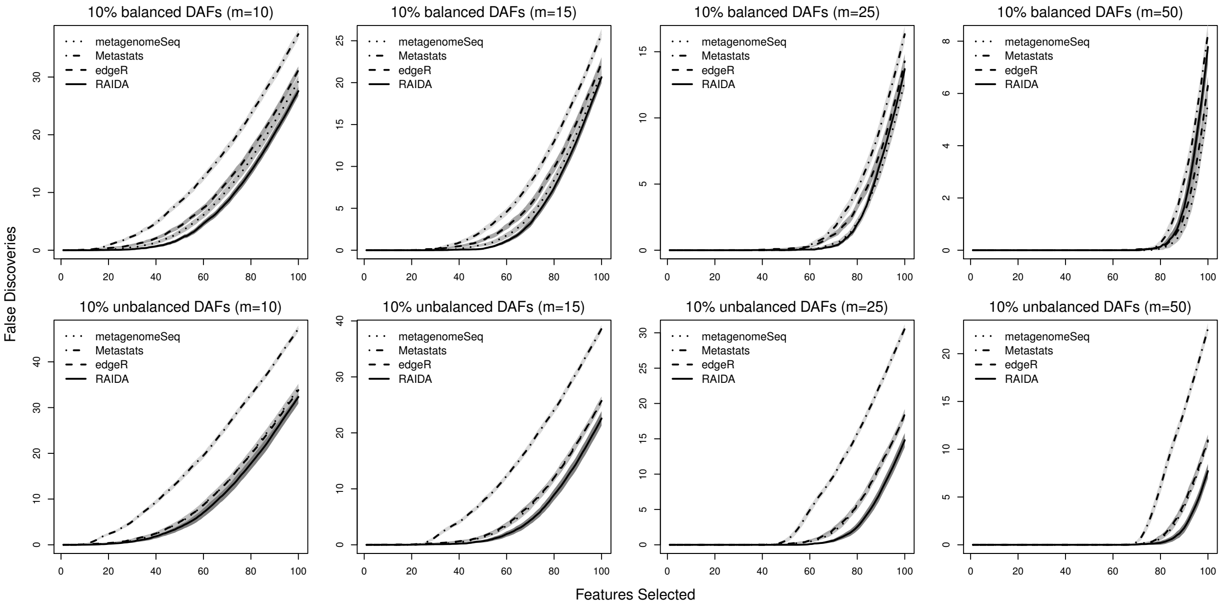

Figure S4: Mean false discovery plots for 10% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The horizontal axis is the number of features in ascending order of p-value, and the vertical axis is the number of falsely identified features. The shades around the lines are 95% confidence bands.

4.2 20% balanced and unbalanced DAFs

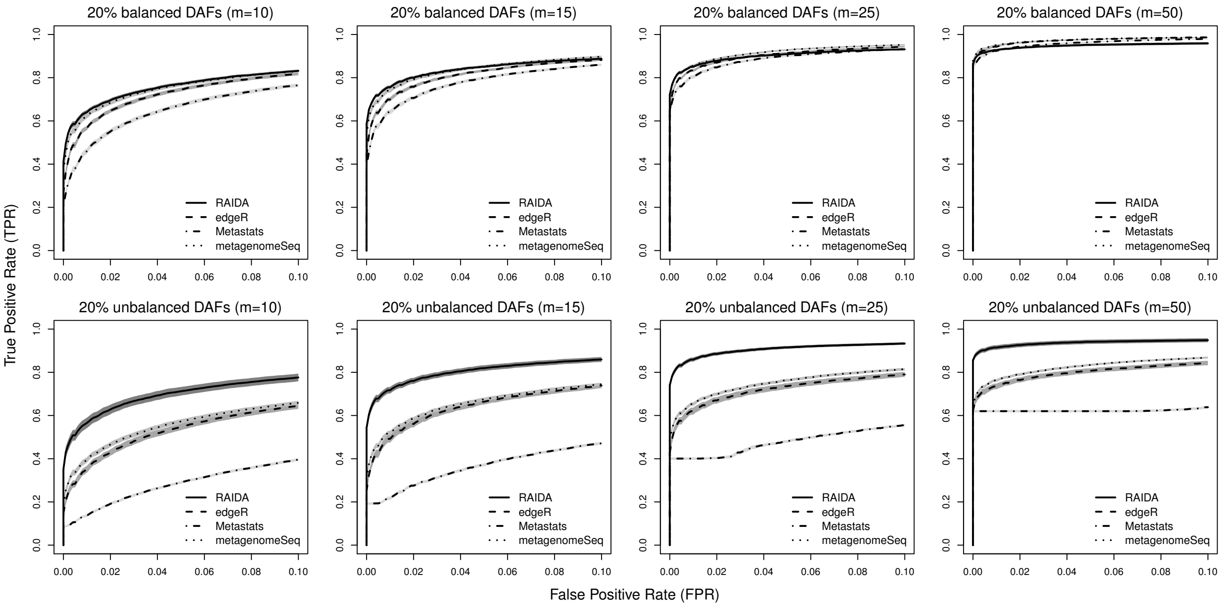

Figure S5: Partial of mean ROC curves for 20% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The shades around the lines are 95% confidence bands.

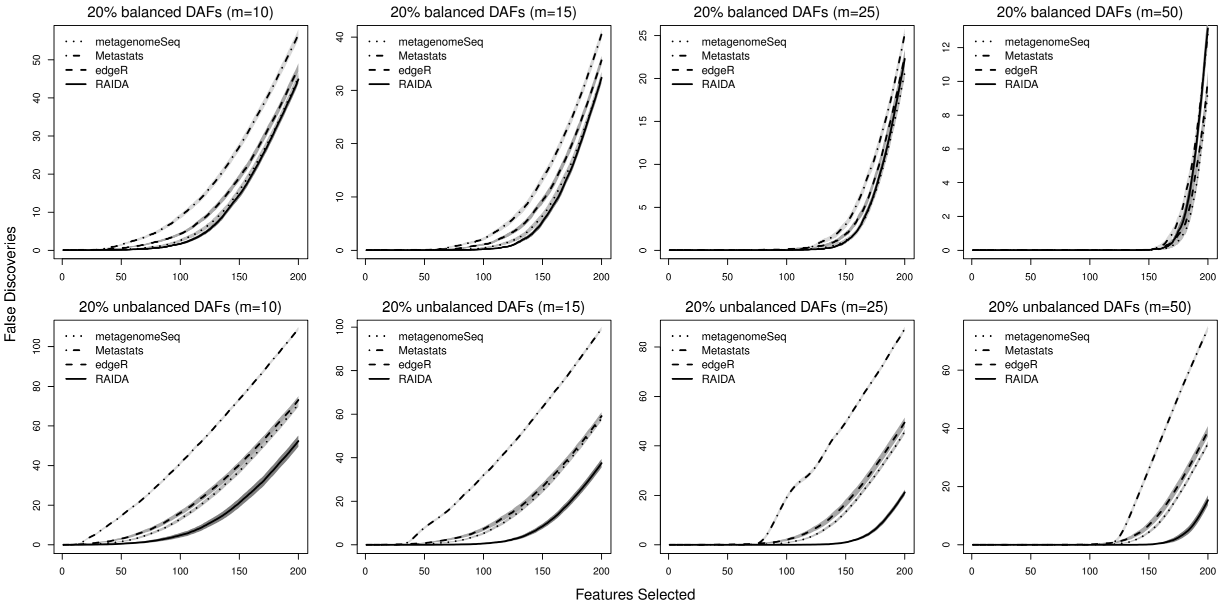

Figure S6: Mean false discovery plots for 20% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The horizontal axis is the number of features in ascending order of p-value, and the vertical axis is the number of falsely identified features. The shades around the lines are 95% confidence bands.

4.3 30% balanced and unbalanced DAFs

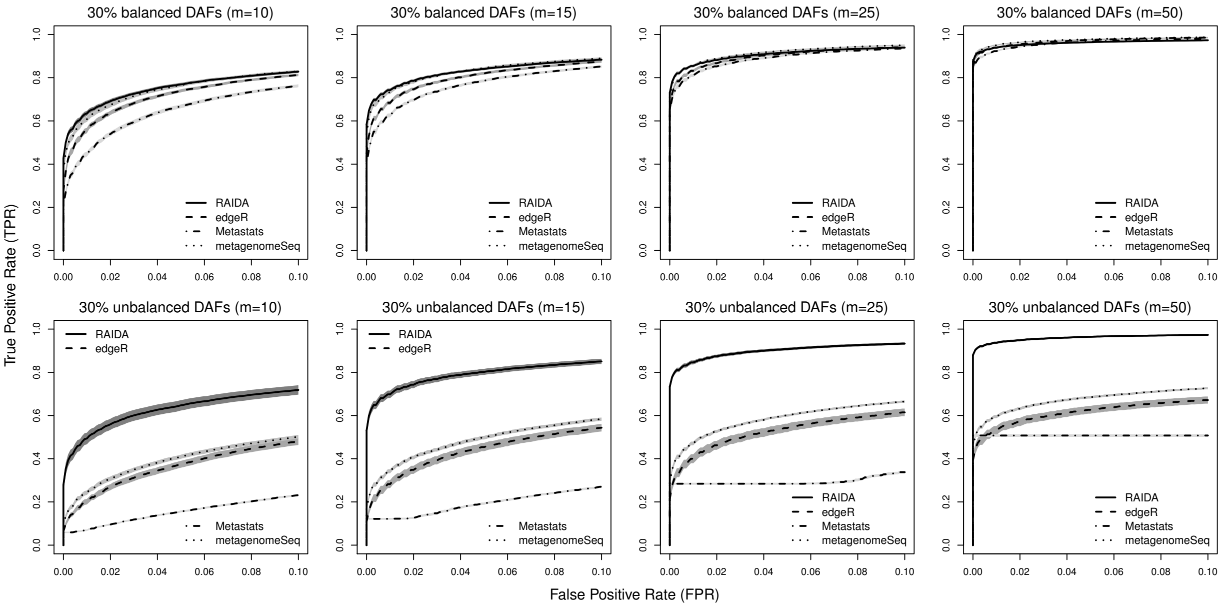

Figure S7: Partial of mean ROC curves for 30% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The shades around the lines are 95% confidence bands.

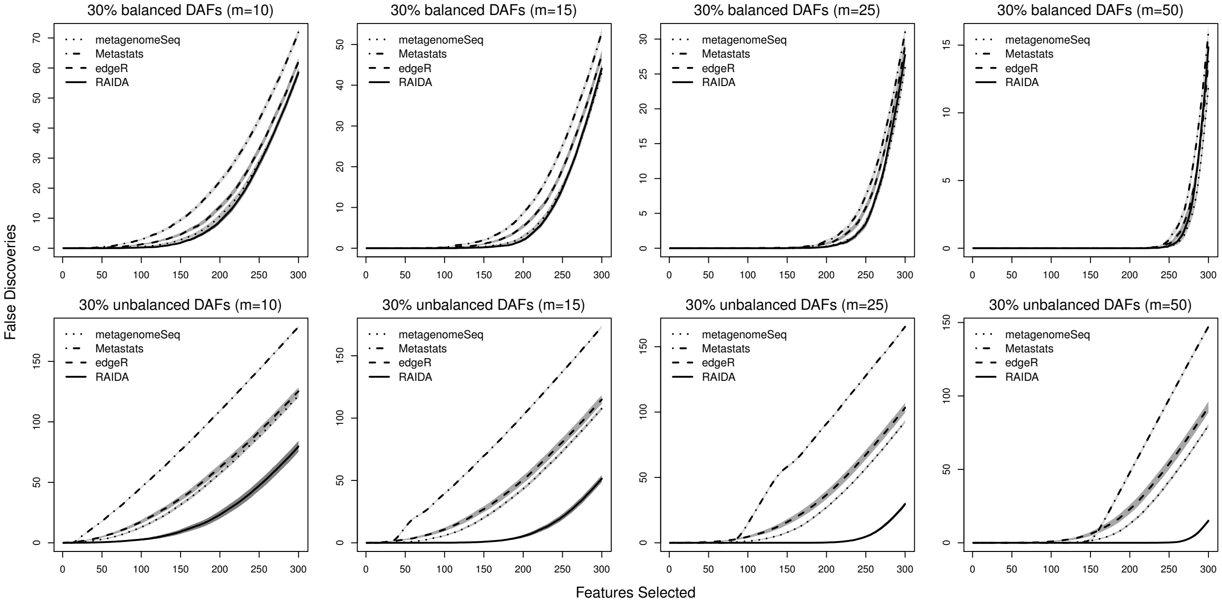

Figure S8: Mean false discovery plots for 30% of DAFs in the balanced and unbalanced conditions with different numbers of samples: $m = 10, 15, 25$ and 50, based on 100 simulations with 1000 features. The horizontal axis is the number of features in ascending order of p-value, and the vertical axis is the number of falsely identified features. The shades around the lines are 95% confidence bands.

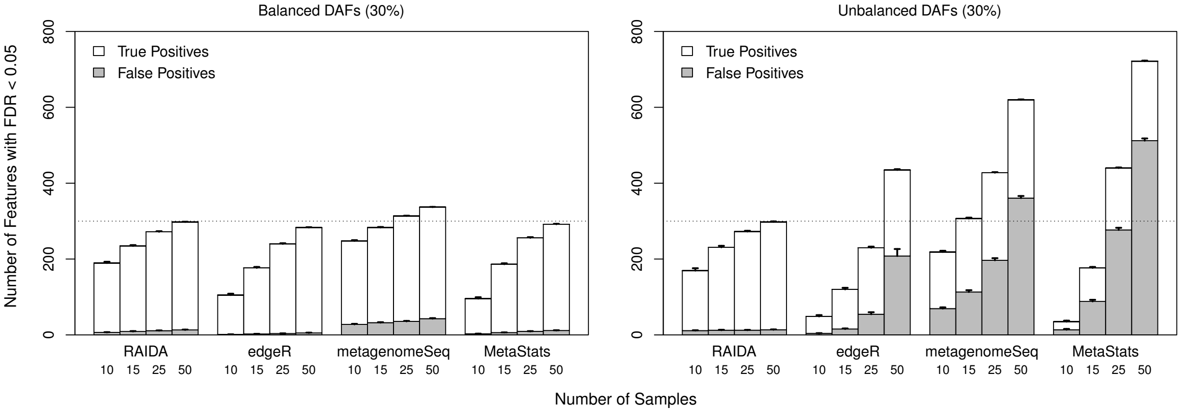

Figure S9: Mean true and false positives plots for 30% of DAFs in the balanced and unbalanced conditions at various numbers of samples, based on 100 simulations with 1000 features. Each bar represents the total number of features that are identified as statistically significant at FDR < 0.05. The white segments are the number of true positives and the gray segments are false positives. The error bars are at a significance level of 0.05. The dashed lines represent the number of DAFs designed for each situation, i.e., there are 300 DAFs out of 1000 features.

4.4 Computation time

A test dataset containing 30 samples (15 samples per each condition) of 1,000 features was used on a desktop with a 3.5 GHz CPU and 16 GB of memory to measure computation time for the tools used in the simulation study. The results are given in Table S3.

Table S3: Comparison of computation time in seconds for the tools used in the simulation study.

edgeR	metagenomeSeq	Metastats	RAIDA
0.4	0.5	149.2	31.6

5 Real data: type II diabetes

5.1 Preprocessing

Table S4: Run numbers for the selected samples for type II diabetes data in NCBI.

Condition	Run number
Controls	SRR341625, SRR341627, SRR341646, SRR341648, SRR341652, SRR341699, SRR341700, SRR341702, SRR341705, SRR341706, SRR341707, SRR341712, SRR341720, SRR341724, SRR341725
Diabetics	SRR341590, SRR341591, SRR341592, SRR341593, SRR341595, SRR341596, SRR341598, SRR341607, SRR341608, SRR341610, SRR341611, SRR341613, SRR341614, SRR341615, SRR341662

We downloaded the files in .sra format from the NCBI archive site, `ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR341/SRR341xxx/SRR341xxx.sra`, where xxx is the last three digits of the Run numbers in Table S4. We converted them into .fasta files using an NCBI SRA Toolkit, fastq-dump. We then split paired-end reads for each sample and rendered sequencing alignments with the forward reads against the bacterial genome references in NCBI using BLASTN since the pairing information is unrelated to performing the alignment with BLASTN. We use 1e-10 for the *e-value* as a cutoff in BLASTN. To parse the BLAST outputs, we used TAEC with the similarity matrix for the read length of 100 bp to get the composition of samples at the species level. We didn't take the bootstrapping option in TAEC because of the colossal sizes of the BLAST outputs: the average size is 6.41 GB, ranging from 2.85 GB to 14.48 GB. Instead, we inspected the numbers of uniquely assigned reads and the similarity factors for similar genomes to identify suspicious species.

5.2 Results

Raw and adjusted p-values for the species whose raw p-values are less than 0.05 are given in Table S5.

Table S5: A list of the species whose p-values for the mean difference in log ratio between the fecal DNA samples of the diabetics and the controls are less than 0.05. The adjusted p-values are computed with the BH method.

species	p-value	adj. p-value
Clostridium botulinum	0.0001	0.0202
Clostridium cellulovorans	0.0004	0.0241
Clostridium beijerinckii	0.0019	0.0876
Parabacteroides distasonis	0.0077	0.2101
Fusobacterium nucleatum	0.0122	0.2348
Odoribacter splanchnicus	0.0137	0.2348
Klebsiella pneumoniae	0.0208	0.2584
Alistipes finegoldii	0.0259	0.2734
Bacteroides thetaiotaomicron	0.0281	0.2746
Bacteroides fragilis	0.0331	0.2862
Tannerella forsythia	0.0428	0.3417
Bacteroides salanitronis	0.0478	0.3417

6 Comparison of results obtained by different values of ϵ

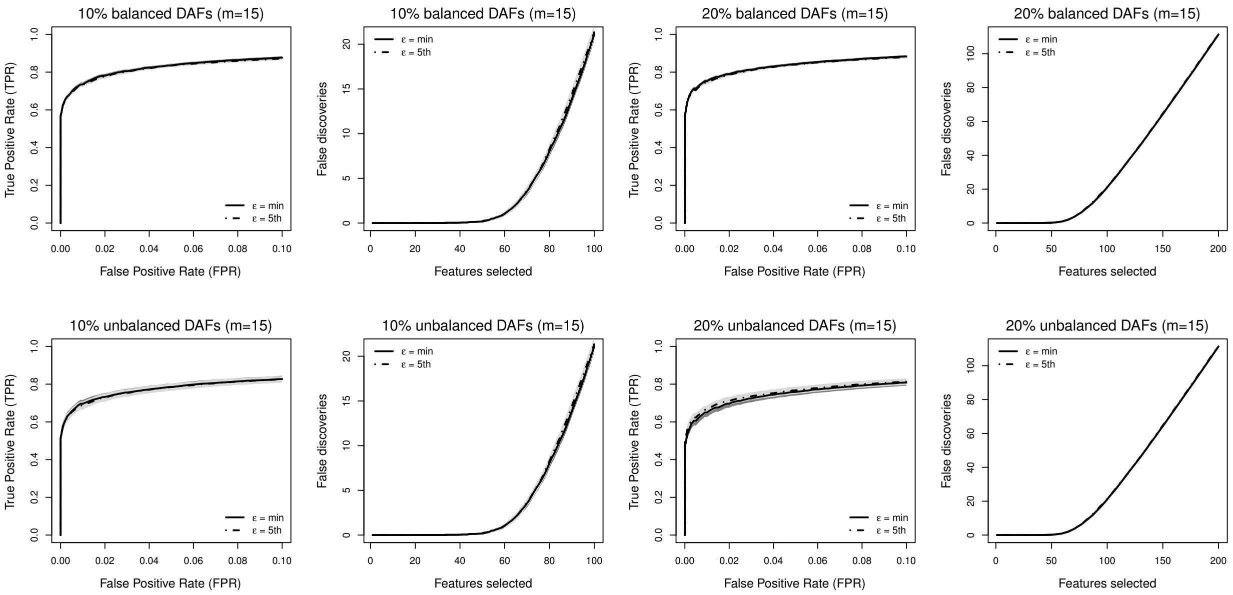

Figure S10: Mean partial ROC curves and mean false discovery plots for 10% and 20% DAFs in the balanced and unbalanced conditions with the sample size of $m = 15$ obtained by different values of ϵ: the minimum non-zero ratio and the smallest number in the 5th percentile. The results are based on 100 simulations with 1000 features. The shades around the lines are 95% confidence bands.