Shankar *et al. BMC Bioinformatics* (2015) 16:31
DOI 10.1186/s12859-015-0467-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses

Jyoti Shankar[1*], Sebastian Szpakowski[1], Norma V Solis[2], Stephanie Mounaud[1], Hong Liu[2], Liliana Losada[1], William C Nierman[1] and Scott G Filler[2,3]

Abstract

Background: Microbiome studies incorporate next-generation sequencing to obtain profiles of microbial communities. Data generated from these experiments are high-dimensional with a rich correlation structure but modest sample sizes. A statistical model that utilizes these microbiome profiles to explain a clinical or biological endpoint needs to tackle high-dimensionality resulting from the very large space of variable configurations. Ensemble models are a class of approaches that can address high-dimensionality by aggregating information across large model spaces. Although such models are popular in fields as diverse as economics and genetics, their performance on microbiome data has been largely unexplored.

Results: We developed a simulation framework that accurately captures the constraints of experimental microbiome data. Using this setup, we systematically evaluated a selection of both frequentist and Bayesian regression modeling ensembles. These are represented by variants of stability selection in conjunction with elastic net and spike-and-slab Bayesian model averaging (BMA), respectively. BMA ensembles that explore a larger space of models relative to stability selection variants performed better and had lower variability across simulations. However, stability selection ensembles were able to match the performance of BMA in scenarios of low sparsity where several variables had large regression coefficients.

Conclusions: Given a microbiome dataset of interest, we present a methodology to generate simulated data that closely mimics its characteristics in a manner that enables meaningful evaluation of analytical strategies. Our evaluation demonstrates that the largest ensembles yield the strongest performance on microbiome data with modest sample sizes and high-dimensional measurements. We also demonstrate the ability of these ensembles to identify microbiome signatures that are associated with opportunistic *Candida albicans* colonization during antibiotic exposure. As the focus of microbiome research evolves from pilot to translational studies, we anticipate that our strategy will aid investigators in making evaluation-based decisions for selecting appropriate analytical methods.

Keywords: Microbiome analysis, Sparse ensemble-based regression, Simulation design, Stability selection, Bayesian model averaging, Evaluation

Background

Biological experiments that incorporate extensive next-generation sequencing (NGS) to profile microbial communities are becoming increasingly common in the field of medical microbiology and elsewhere [1-3]. In a substantial number of these studies, the primary objective is to establish associations between the microbial community and specific biological traits or endpoints of clinical relevance. Analytical methods to test these associations follow three general directions. The first explains microbial abundance profiles in terms of biological and clinical variables [4-11]. The second utilizes these microbial profiles to classify samples into biological or clinical categories (e.g., human body sites, health and disease states, subject identities, etc.) [12-15]. A third explains a clinical or biological endpoint in terms of microbial

*Correspondence: jyoti.shankar@gmail.com
[1] J. Craig Venter Institute, 9704, Medical Center Drive, Rockville, 20850 Maryland, US
Full list of author information is available at the end of the article

community profiles in combination with other biological covariables [16,17]. The first two directions have formed the basis for most of the analytical work in this field. In this work, we explore the third direction.

Specifically, we used a mouse model to investigate the risk factors during antibiotic therapy for opportunistic, gastrointestinal (GI) colonization with *Candida albicans*, a commensal fungus in the human GI tract (Figure 1 and Table 1). *C. albicans* does not normally colonize mice unless the resident GI flora is perturbed by antibiotics. Therefore, we exposed mice to a single challenge with *C. albicans* 7 days after initiating antibiotic treatments that spanned a total of 21 days. Our initial observations indicated that some antibiotics induced a persistently higher level of *C. albicans* colonization compared to others. We sought further insights into these differential patterns of colonization by identifying the underlying antibiotic-induced perturbations in the bacterial and fungal GI microbiome, and the host immune factors. Hence, we employed a regression framework to explain the level of *C. albicans* colonization using microbiome variables, immune factors and experimental conditions as covariables.

We faced unique challenges in performing these analyses. Our studies utilized animal models with tens of samples (n). However, the total number of variables (p) arising from NGS measurements of the microbiome, host immune factors and other biological exposures was at least an order of magnitude larger than the number of samples ($p >> n$). In this scenario, a model estimated using univariate or non-regularized regression modeling strategies [6,8] is likely to overfit the data and yield findings that would not generalize well on data from newer studies [18]. Moreover, this model does not take into account the covariance structure of high-dimensional data, making the naïve assumption of orthogonality of covariables. In addition, such a model is only one among the 2^p possible model configurations that could explain the data [19,20]. We therefore needed a modeling strategy which a) incorporates penalized regression to avoid overfitting, b) addresses multicollinearity, c) efficiently searches through a large number of model configurations prior to settling on the most likely model specification, and d) estimates the model parameters with low variability in the face of modest sample sizes.

Ensemble models are a class of approaches that elegantly satisfy all of the above requirements by merging information on the explanatory variables from a large collection of models to generate a ranked list of influential variables ordered by their importance [20]. These ensembles estimate robust models with stable solutions by utilizing the inherent variability in either the sample space through bootstrap methods [21,22] or the model space using Bayesian methods [23] and Markov chain Monte Carlo (MCMC) algorithms [24].

Random forests [25], a non-linear ensemble method, has been widely used for classification and prediction tasks in microbiome research [12,14]. However, random forests are not directly usable in a problem domain such as ours that requires the estimation of a model in a linear regression framework for a continuous clinical endpoint of interest. Investigators in the field have adopted penalized regression methods [4,5,14-16] that are able to effectively model multivariable covariances within the microbiome. Nevertheless, these methods have been employed in a non-ensemble setting which could potentially lead to instability in model findings.

Only a few research groups have incorporated ensemble linear regression modeling in a problem domain similar to ours [17]. Furthermore, there has been no assessment of how the characteristics of linear regression ensembles affect model performance on realistic microbiome data. We address this gap in the microbiome literature by a) rigorously evaluating the performance of

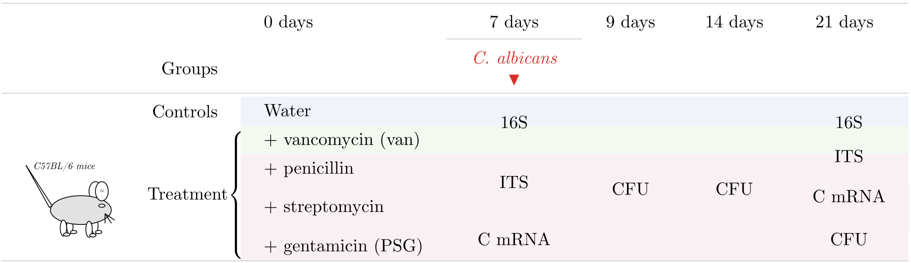


Figure 1 Experimental design of the mouse microbiome study. Mice were divided into a control group and two treatment groups. The treatment groups received either vancomycin (van) or a combination of penicillin, streptomycin and gentamicin (PSG). After 7 days, some of the mice were administered a *C. albicans* challenge. Terminal ileum samples were collected on days 7 and 21 and fecal pellet samples on days 7, 9, 14 and 21. *C. albicans* colonization level was measured by counting the Colony forming units (CFU) following quantitative culture of the fecal pellets. Cytokine mRNA expression (C mRNA) was measured in the terminal ileum samples. Bacterial 16S and Fungal ITS were amplified from both the terminal ileum sections and the fecal pellets. These amplicons were sequenced with 454 pyrosequencing. The YAP bioinformatic workflow [54] was used to obtain genus-level taxonomic assignments for the bacterial and fungal sequences.

Table 1 Characteristics of data from the mouse microbiome study

Measurement	Original variable	Pre-processing	Dimension	Role in models
16S	Sequence counts	*log* (Relative abundances)	344	Independent
ITS	Sequence counts	*log* (Relative abundances)	109	Independent
C mRNA	Expression levels	*log* (GAPDH normalized expression levels)	6	Independent
CFU	Colony counts	$log\left(\frac{\text{Colony forming units (CFU)}}{\text{grams of fecal matter}}\right)$	1	Dependent (Response)
Antibiotic treatment ± *C. albicans* challenge + Length of treatment	Categorical	Dummy encoding	5 levels	Independent

C. albicans colonization level was measured by counting the Colony forming units (CFU) following quantitative culture of the fecal pellets. Cytokine mRNA expression (C mRNA) was measured in the terminal ileum samples. Bacterial 16S and Fungal ITS were amplified from both the terminal ileum sections and the fecal pellets. These amplicons were then sequenced by 454 pyrosequencing and taxonomically classified with a tailored bioinformatic pipeline [54].

a selection of frequentist and Bayesian ensemble regression modeling approaches, and b) introducing a simulation and evaluation framework that can be constructed to closely mimic the characteristics of any experimental microbiome dataset. Using this framework and a suite of metrics, we determine how various characteristics of ensembles influence model performance. We also demonstrate the application of these ensembles to our experimental mouse data. Results from our evaluation demonstrate that ensembles capable of exploring larger model spaces perform better with lower variability.

Methods

Simulation design

We generated data for our simulation study using the linear model,

$$Y = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where Y is the response variable, $\boldsymbol{X}$ is the design matrix with n samples and p variables, $\boldsymbol{\beta}$ is the p-dimensional vector of regression coefficients and the random error $\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 \cdot \mathrm{I}_n\right)$ where I_n is the identity matrix. We describe the construction of the components $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ below.

Design matrices

We constructed two design matrices based on the experimental mouse gut bacterial and fungal microbiome profiles, respectively. We first eliminated the rows and columns which had zero counts across all samples; n and p refer to the total number of rows and columns that remained. In the bacterial matrix, $n = 30$ and $p = 192$ while in the fungal matrix, $n = 36$ and $p = 47$. We then constructed the design matrix $\boldsymbol{X}$ by log-transforming the relative abundances of genera, t'_{ij} where $i \in \{1, 2, \ldots n\}$ and $j \in \{1, 2, \ldots, p\}$.

$$\phi_{ij} = \left(\frac{t_{ij} + 1}{\sum_{j=1}^{p} t_{ij} + 1}\right), t'_{ij} = log(\phi_{ij}), \tag{2}$$

where ϕ_{ij} and t_{ij} denote the relative abundance and the sequence count of bacterial or fungal genus, j in sample i, respectively. We added a smoothing constant of 1 to t_{ij} to ensure non-zero proportions for all genera. The log-transformation ensures that all variables have a similar dynamic range across all the samples [26]. p denotes the total number of genera.

Since $\sum_{j=1}^{p} \phi_{ij} = 1$ for every i, the relative abundance of one of the genera is fully specified by the others. To address this redundancy, we selected one of the genera as a reference, and dropped it from the design matrix $\boldsymbol{X}$. We chose the reference genus, j_{ref} to be the one with the greatest abundance across all the samples. i.e.,

$$j_{\text{ref}} = \underset{j \in \{1,2,\ldots p\}}{\operatorname{argmax}} \sum_{i=1}^{n} t_{ij}. \tag{3}$$

Influential variables

Microbiome sequence data describes an entire community of organisms. It is not known whether this community consists of a large or a small subset of influential microbes. In addition, their degree of association with the response is also unknown. We addressed this uncertainty using the following strategies: a) considering a range of sparsity settings (% of variables specified as influential $\in \{2\%, 3\%, 4\%\}$), and b) sampling regression coefficients from either a bimodal distribution, $\beta = \pm 1$ or a uniform distribution, $\beta \in \mathcal{U}$.

We denote the set of specified influential variables as $\mathcal{V}_i$. We randomly selected the indices of these variables and divided them into two halves $\mathcal{V}_i^1$ and $\mathcal{V}_i^2$. We used two alternative schemes for assigning regression coefficients $\boldsymbol{\beta}$ to the selected variables in $\mathcal{V}_i^1$ and $\mathcal{V}_i^2$, following the approach in Meinshausen et al. [22].

1. $\beta = \pm 1$: $\forall \nu \in \mathcal{V}_i^1, \beta_\nu = 1$; $\forall \nu \in \mathcal{V}_i^2, \beta_\nu = -1$.
2. $\beta \in \mathcal{U}$: $\forall \nu \in \mathcal{V}_i^1, \beta_\nu \sim \mathcal{U}(0.5, 1.0)$; $\forall \nu \in \mathcal{V}_i^2, \beta_\nu \sim \mathcal{U}(-1.0, -0.5)$.

Signal to noise ratio

The signal to noise ratio (SNR) is defined as [27]

$$\text{SNR} = \frac{||\boldsymbol{X\beta}||_2}{\sqrt{n}\sigma}, \tag{4}$$

where $||\boldsymbol{X\beta}||_2$ is the l_2 norm of the predicted response. We computed the SNRs in our experimental mouse data by estimating an elastic net model [28]. The bacterial data had an SNR of 4.6, while the fungal data had an SNR of 5.2. In our simulation, we applied three different SNRs of 0.25, 4.60 and 16.00, that correspond to high, medium, and low noise settings, respectively. For a given SNR, we sampled the random error $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where $i \in \{1, 2, \ldots n\}$ and where σ is chosen as:

$$\sigma = \frac{||\boldsymbol{X\beta}||_2}{\sqrt{n} \cdot \text{SNR}}. \tag{5}$$

It is easier to recover the influential variables under a higher SNR since the response has a larger association with the variables relative to the random error term.

For each combination of $\boldsymbol{X}$, $\boldsymbol{\beta}$ and SNR, we performed a total of 130 simulations and report results aggregated over these runs, making our findings robust to variations. We conducted our evaluation and analysis in the R language for statistical computing [29] and have released our code at *GitHub* [30].

Modeling approaches

In our evaluation, elastic net with cross-validation (ENC) was the baseline non-ensemble penalized regression method. We chose the remaining frequentist and Bayesian ensemble modeling approaches with the objective of determining how the size of the ensemble, construction of the model space, and the choice of regularization parameters within the ensemble influenced performance. We summarize these approaches in Table 2 and describe them in detail below.

Elastic net with cross-validation (ENC)

The ENC procedure [28] is a specific instance of a penalized regression model [31,32].

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} |y_i - X_i\boldsymbol{\beta}|^2 + \lambda \left(\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 \right) \right\}, \tag{6}$$

where $\hat{\boldsymbol{\beta}}$ is the solution to the estimation problem, the penalty terms, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\boldsymbol{\beta}_i|$ and $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{i=1}^{p} |\boldsymbol{\beta}_i|^2}$ are the l_1 and l_2 norms of the coefficient vector $\boldsymbol{\beta}$, $\sum_{i=1}^{n} |y_i - X_i\boldsymbol{\beta}|^2$ is the residual sum of squares (RSS), λ is the tuning parameter that penalizes the RSS by the size of the regression coefficients and α is a tuning parameter that balances the l_1 and l_2 penalties. An α of 1 promotes sparsity in the model, while an α of 0 ensures that correlated variables are assigned similar regression coefficients. An optimal value of α finds a balance between the two penalties. We employed leave-one-out cross-validation to determine the optimal combination of α and λ ($\lambda_{optimal} = \lambda_{ENC}$) that minimized the mean squared error (MSE) of the model [31]. We used the implementation of ENC available in the *glmnet* R package [32].

Frequentist ensembles

We evaluated frequentist ensembles that were variations on the ensemble strategy originally described by Meinshausen and Bühlmann (MB) [22]. In their work, MB proposed a method to improve stability of variable selection within the least absolute shrinkage and selection operator (LASSO) approach [33]. Since ENC is a

Table 2 Summary of the modeling approaches included in the evaluation

Model	**Ensemble characteristics**		**Output**	**Paradigm**	**R Package**
	Tuning parameter	**Model space construction**			
ENC	λ_{ENC}	None	Influential variables	Frequentist (l_1, l_2 penalties)	*quadrupen, glmnet*
PS	λ_{MB}	Subsampling	Influential variables		
LS	λ_{ENC}		Inclusion probabilities		
SS	Λ		Inclusion probabilities		
PR	λ_{MB}	Resampling	Influential variables		
LR	λ_{ENC}		Inclusion probabilities		
SR	Λ		Inclusion probabilities		
BMA	$EMS = 1$	MCMC	Inclusion probabilities	Bayesian (Spike & slab prior)	*BoomSpikeSlab*
BMAC	EMS_{CV}		Inclusion probabilities		

ENC: The baseline penalized regression model. Elastic net with $\lambda_{optimal} = \lambda_{ENC}$ derived from cross-validation (CV), **Ensembles based on 100 subsamples: PS:** Meinshausen & Bühlmann's algorithm with a single $\lambda_{optimal} = \lambda_{MB}$ selected to minimize the expected number of false positives, **LS:** Single $\lambda_{optimal} = \lambda_{ENC}$ with no variable selection, **SS:** Stability selection across the entire 100 $\lambda \in \Lambda$ grid with no variable selection, **Ensembles based on 100 resamples: PR, LR, SR:** Identical to PS, PR and LR, respectively, with model space constructed through resampling. **BMA:** Bayesian model averaging with expected model size (*EMS*) = 1, **BMAC:** BMA with EMS determined by CV (EMS_{CV}).

generalization of the LASSO approach and our baseline, we replaced LASSO with ENC in MB's stability selection procedure.

Ensembles based on subsamples (PS, LS, SS, SSW)
Within MB's stability selection procedure, we first generated 100 subsamples of the data. From each subsample, we estimated regression coefficients $\boldsymbol{\beta}$ on a grid of regularization parameters $\lambda \in \Lambda$. Λ consisted of 100 λ values obtained using the Least angle regression (LARS) algorithm within ENC [31,34] implemented in the *glmnet* R package [32]. Given the set of selected variables at any λ, $S(\lambda) = \{j : \beta_j(\lambda) \neq 0\}$, the inclusion probability $P(j \in S(\lambda))$ of any variable $j \in \{1, 2, \ldots p\}$ was computed as

$$P(j \in S(\lambda)) = \frac{1}{B} \sum_{b=1}^{B} 1_{j \in S_{\lambda}(b)}, \tag{7}$$

where b is the index of any given subsample, $B = 100$ is the total number of subsamples and $S_{\lambda}(b)$ is the set of selected variables at regularization parameter λ in subsample b. To minimize the expected number of false positives within the selected set of variables, $S(\lambda)$, MB developed an algorithm to select both the optimal regularization parameter $\lambda_{optimal} = \lambda_{MB}$ and a stable set of variables, $\hat{S}^{stable} \subseteq S(\lambda_{MB})$ [22]. We adopted an implementation of MB's algorithm from the *quadrupen* R package [35]. Since MB's algorithm was designed to reduce false positives and employed subsampling to construct the model space, we called this modeling approach, PS. In addition to PS, we evaluated ensemble variants which enabled us to determine how performance was influenced by the choice of the regularization parameter $\lambda_{optimal}$, size and nature of the model space, and the computation strategy for inclusion probabilities.

In the first variant of PS, we chose the single regularization parameter $\lambda_{optimal} = \lambda_{ENC}$, specified by the ENC procedure in *glmnet* [32]. Unlike PS that performed variable selection at λ_{MB}, we computed the inclusion probabilities of the p variables across the 100 subsamples at λ_{ENC} using Equation 7. We termed this variant LS since it operated at a single $\lambda_{optimal}$ and employed subsampling to construct the model space.

In the next variant, we expanded the size of the model space by computing the average inclusion probabilities, I_j, of the variables across the 100 regularization parameters $\lambda \in \Lambda$,

$$I_j = \frac{1}{100} \cdot \sum_{i=1}^{100} P(j \in S(\lambda_i)), \tag{8}$$

where $j \in \{1, 2, \ldots p\}$. The approach assigned high inclusion probabilities to the stable variables which were most consistently selected across the Λ grid and employed subsampling to construct the model space. Hence we termed this variant SS.

We also considered an alternative definition of inclusion probability, $I_j^w, j \in \{1, 2, \ldots, p\}$ based on a suggestion from our referee.

$$I_j^w = \sum_{i=1}^{100} w_{\lambda_i} \cdot P(j \in S(\lambda_i)), \tag{9}$$

where w_{λ_i}, the weight assigned to λ_i, is proportional to the average number of non-zero variables at λ_i

$$w_{\lambda_i} = \frac{\sum_{b=1}^{B} 1_{j \in S_{\lambda_i}(b)}}{\sum_{i=1}^{100} \sum_{b=1}^{B} 1_{j \in S_{\lambda_i}(b)}}, \tag{10}$$

where b is the index of the subsamples and $B = 100$ is the total number of subsamples. We termed this variant SSW.

Ensembles based on resamples (PR, LR, SR, SRW)
Since MB [22,27] recommend the use of resampling as an equivalent alternative to subsampling in the context of stability selection, we varied the nature of model space construction within the frequentist ensembles by adopting bootstrap resampling as an alternative to subsampling with size $\lfloor \frac{n}{2} \rfloor$. Subsampling without replacement and sampling at random with replacement (bootstrap resampling) are examples of schemes within a family of *exchangeably weighted bootstrap schemes* [36,37]. The equivalence of the two bootstrapping schemes has been described in [38] and [39] and demonstrated through theoretical results and extensive simulations of non-linear regression [40]. In addition, the two schemes have very similar statistical properties [41]. It has also been independently shown that bootstrap resampling improves stability of the LASSO procedure, which is a special case of ENC [42].

We implemented the resampling alternatives to PS, LS, SS and SSW, and termed them PR, LR, SR and SRW, respectively.

Bayesian ensembles

In addition to the frequentist ensembles, we evaluated Bayesian ensembles that further expanded the size of the model space explored. These larger ensembles were based on spike-and-slab Bayesian model averaging (BMA) [24] implemented in the *BoomSpikeSlab* R package [43]. Given a set of multicollinear variables, spike-and-slab BMA assigns higher inclusion probabilities to influential variables highly associated with the response. Strongly correlated variables that are equally associated with the response are assigned similar inclusion probabilities. Thus the BMA framework is conceptually similar to our baseline frequentist penalized regression approach, ENC, and its ensemble variants.

The spike-and-slab prior consists of two components and is described in detail in [44,45]. The spike component of the prior is modeled by a Bernoulli distribution for each variable and is described by

$$\boldsymbol{\gamma} \sim \prod_{i=1}^{p} \pi_i^{(\gamma_i)}(1-\pi_i)^{(1-\gamma_i)}, \tag{11}$$

where $\boldsymbol{\gamma}$ is a vector of binary indicators, $\gamma_i; i \in \{1, 2 \ldots p\}$. If $\gamma_i = 1$, the i^{th} variable is selected as influential. If $\gamma_i = 0$, the i^{th} variable is not influential and $\beta_i = 0$. π_i is the probability of the i^{th} variable being influential and computed as $\pi_i = \frac{k}{p}$. k is the expected model size and specifies the number of variable coefficients expected to be non-zero.

Conditional on the set of influential variables $\{i : \gamma_i = 1\}$, the slab component of the prior distribution models the regression coefficients $\boldsymbol{\beta}$ for these variables. The distribution is a variant of the Zellner's g-prior [46] and is given by

$$\boldsymbol{\beta} | \left(\sigma^2, \boldsymbol{\gamma}\right) \sim \mathcal{N}\left(\boldsymbol{b}, \boldsymbol{F_\gamma}^{-1}\right), \tag{12}$$

where $\boldsymbol{b}$ denotes the prior expectation of $\boldsymbol{\beta}$ and is set to $\mathbf{0}$ in our experiments. $\boldsymbol{F_\gamma}$ is the sub-matrix that corresponds to the influential variables $\{i : \gamma_i = 1\}$ within the full prior Fisher information matrix,

$$\boldsymbol{F} = \frac{g}{n} \cdot \frac{\boldsymbol{X^T X}}{\sigma^2}. \tag{13}$$

In the event of multicollinearity, $\boldsymbol{F}$ may not always be positive definite. Therefore, the BMA implementation within *BoomSpikeSlab* ensures a proper posterior distribution by linearly interpolating $\boldsymbol{X^T X}$ with its diagonal to obtain the smoothed Fisher information matrix [44,45],

$$\boldsymbol{F}_{smooth} = \frac{g}{n} \cdot \frac{w \times \boldsymbol{X^T X} + (1-w) \times \mathrm{diag}(\boldsymbol{X^T X})}{\sigma^2}, \tag{14}$$

where g is set to 1 and $w = 0.5$. σ^2 is the variance of the random error $\boldsymbol{\epsilon}$ in the regression model and distributed as

$$\frac{1}{\sigma^2} | \boldsymbol{\gamma} \sim \Gamma\left(\frac{\nu}{2}, \frac{ss}{2}\right), \tag{15}$$

where $\Gamma\left(\frac{\nu}{2}, \frac{ss}{2}\right)$ represents a gamma distribution with mean $\frac{\nu}{ss}$ and variance $\frac{\nu}{ss^2}$. We retained the default values of $\nu = 0.01$ and $ss = 0.5 \cdot s_y^2$ from the *BoomSpikeSlab* R package [43]. s_y^2 is the standard deviation of the response. We set the expected model size, k, to the default value of 1 in the variant termed BMA. For the variant termed BMAC, we estimated k with five-fold cross-validation to minimize the MSE. This enabled the BMAC variant to adapt to the inherent sparsity setting in the data.

The BMA procedure incorporates a MCMC algorithm that traverses a very large space of models to estimate the posterior distribution of the regression model parameters [44]. For our datasets, the running means of the selection indicators and the regression coefficients converged well before 10,000 MCMC iterations. Hence, we ran the MCMC algorithm for 10,000 iterations and discarded the initial 1000 iterations as burn-in. We then estimated the variable inclusion probabilities from the remaining 9000 iterations. The inclusion probability for the i^{th} variable is the proportion of iterations (or draws) with non-zero regression coefficients

$$P(\gamma_i = 1) = \frac{1}{D} \sum_{d=1}^{D} 1_{\beta_i \neq 0}(M_d), \tag{16}$$

where $D = 9000$ is the total number of draws of model parameters from the posterior distribution, and M_d indicates the model parameters at iteration d.

Variable selection

In high-dimensional settings, variable selection results in a sparse set of variables which provide the best explanation for the data. ENC performs variable selection by assigning the non-influential regression coefficients to zero. PS and PR directly compute the set of influential variables. For all the other approaches that generate variable inclusion probabilities, we applied a variable selection algorithm. In this algorithm, we first ranked the inclusion probabilities in decreasing order. We then computed the first order lagged differences between the ranked inclusion probabilities, identified the largest gap in this sequence, and selected the variables above this gap as influential. Suppose $\mathrm{IP}_i, i \in \{1, 2, \ldots p\}$ denote the ranked inclusion probabilities, then the index that corresponds to the largest gap in inclusion probabilities is

$$\hat{i} = \underset{i}{\operatorname{argmax}} \{\mathrm{IP}_i - \mathrm{IP}_{i+1}\}, \tag{17}$$

and the set $\hat{\mathcal{V}_i} = \{v_1, v_2, \ldots v_{\hat{i}}\}$ is the chosen set of influential variables. We refer to this strategy as the lagged differences (LD) algorithm.

Evaluation metrics

We evaluated the approaches on the basis of their ability to select the truly influential variables as well as their ability to rank these variables accurately.

Variable selection

ROC curves and AUC score The receiver operating characteristic (ROC) curve is a visual depiction of the variable selection performance of an approach across a range of thresholds on the inclusion probability. We constructed the ROC curve for each approach by plotting true positive rate (TPR) versus false positive rate (FPR) while sweeping across a sequence of 22 inclusion probability thresholds between 0.00 and 1.01 on the logarithmic scale. True positives (TP) are those variables which were correctly identified as influential by an approach. False positives (FP) are those which were incorrectly identified as influential by an approach. TPR and FPR are defined as follows

$$TPR = \frac{TP}{|\mathcal{V}_i|} \; FPR = \frac{FP}{p - |\mathcal{V}_i|}, \tag{18}$$

where p is the total number of variables and $\mathcal{V}_i$ is the set of all influential variables in the simulated data. In constructing the ROC curve for any given approach, variables with inclusion probabilities above a given threshold were counted as influential variables selected at that threshold. We also computed the area under the ROC curve (AUC) score for each of the approaches. The AUC is a single metric that quantifies the performance of any given approach across the complete range of inclusion probability thresholds [47]. Both the ROC curve and the AUC metric allowed us to assess the performance of approaches that generated inclusion probabilities for variables but did not perform variable selection. Therefore, these metrics evaluated model performance independent of any particular algorithm for variable selection.

F-score The F-score is a single metric that measures the accuracy of variable selection with respect to the truly influential variables. It is the harmonic mean of precision and recall and is given by

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP+FP}} \\ \text{Recall} &= \frac{TP}{|\mathcal{V}_i|} \\ \text{F-score} &= 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right). \end{aligned} \tag{19}$$

Variable ranking

We measured the Spearman's rank correlation coefficient between the simulation-assigned ranks $r = (r_1, r_2, \ldots r_p)$ and the observed ranks $\hat{r} = (\hat{r}_1, \hat{r}_2, \ldots \hat{r}_p)$ of the variables. For variable i, r_i was based on the absolute value of its regression coefficient while $\hat{r}_i$ was based on its inclusion probability from a given approach.

The Spearman's rank correlation between r and $\hat{r}$ is specified by

$$\rho = \frac{\sum_{i=1}^{p}(r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sqrt{\sum_{i=1}^{p}(r_i - \bar{r})^2 \sum_{i=1}^{p}(\hat{r}_i - \bar{\hat{r}})^2}}, \tag{20}$$

where $\bar{r}$ and $\bar{\hat{r}}$ are the mean ranks of the true regression coefficients and the inclusion probabilities, respectively.

Results and discussion

Receiver operating characteristic (ROC)

Figure 2 shows the mean ROC curves, computed over 130 simulations, for approaches that assign inclusion probabilities to variables without performing variable selection. Good performance is characterized by high TPR and low FPR over a wide range of inclusion probability thresholds. The diagonal line on the plots represents random variable selection performance. ROC curves below the diagonal indicate poor performance while those above have better than random performance.

Overall, BMA variants consistently outperformed all the other approaches. Performance of all approaches deteriorated in high dimensions and with increasing number of influential variables, demonstrating that all ensembles had uniform difficulty in variable selection within less-sparse settings. Indeed, the setting with the highest percentage (4%) of influential variables and largest regression coefficients ($\beta = \pm 1$) presented the most challenging scenario. The performance gap between BMA and the other approaches decreased in this case.

The relatively strong performance of all approaches in the low-dimensional fungal setting was most likely due to the smaller space of variable configurations. In most scenarios, SS and SR performed better than LS and LR approaches, illustrating the advantage of moving from a single $\lambda_{optimal}$ to a larger model space consisting of a grid $\lambda \in \Lambda$.

At higher SNRs, BMA with cross-validated expected model size (BMAC) performed better than the other approaches. We also note that higher SNRs were uniformly associated with better performance.

Area under the ROC curve (AUC)

Figure 3 shows the variability in AUC across 130 simulations for the approaches that assign inclusion probabilities. The corresponding median AUCs are shown in Additional file 1. An AUC of 1.0 is ideal and corresponds to a scenario where an approach consistently ranks influential variables higher than non-influential variables. In our simulations, the trends in AUC mirrored the patterns in the ROC curves. BMA variants and specifically, BMAC, obtained the highest AUC along with the lowest variability, showing the advantages of exploring the largest model space among all approaches.

While BMA variants had a near perfect AUC in the low-dimensional fungal setting, the AUCs for non-BMA approaches were also higher in this setting. Although SS and SR had AUCs similar to BMA in the fungal setting, they had a much higher variability. The resampling variants, LR and SR, yielded small but consistent improvements in median AUC relative to their subsampling counterparts in the high-dimensional setting.

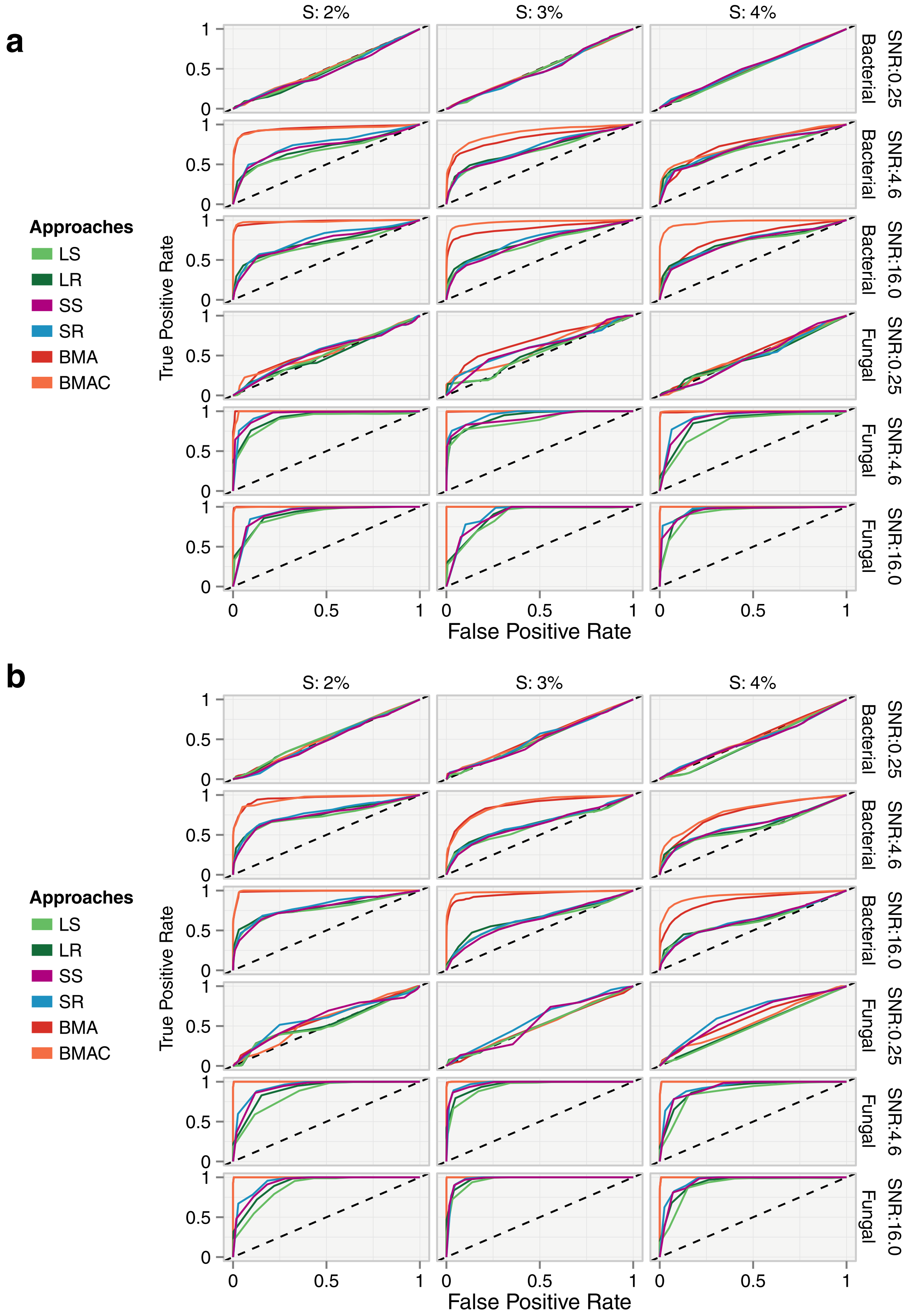


Figure 2 Receiver operating characteristic (ROC) curves. Mean ROC curves across 130 simulations are shown for approaches that do not perform variable selection. An ideal ROC has a high TPR and a low FPR over a wide range of thresholds. *S* denotes the sparsity setting or the proportion of influential variables in the data. The diagonal dashed line represents random variable selection performance. **a**. $\beta = \pm 1$ and **b**. $\beta \in \mathcal{U}$.

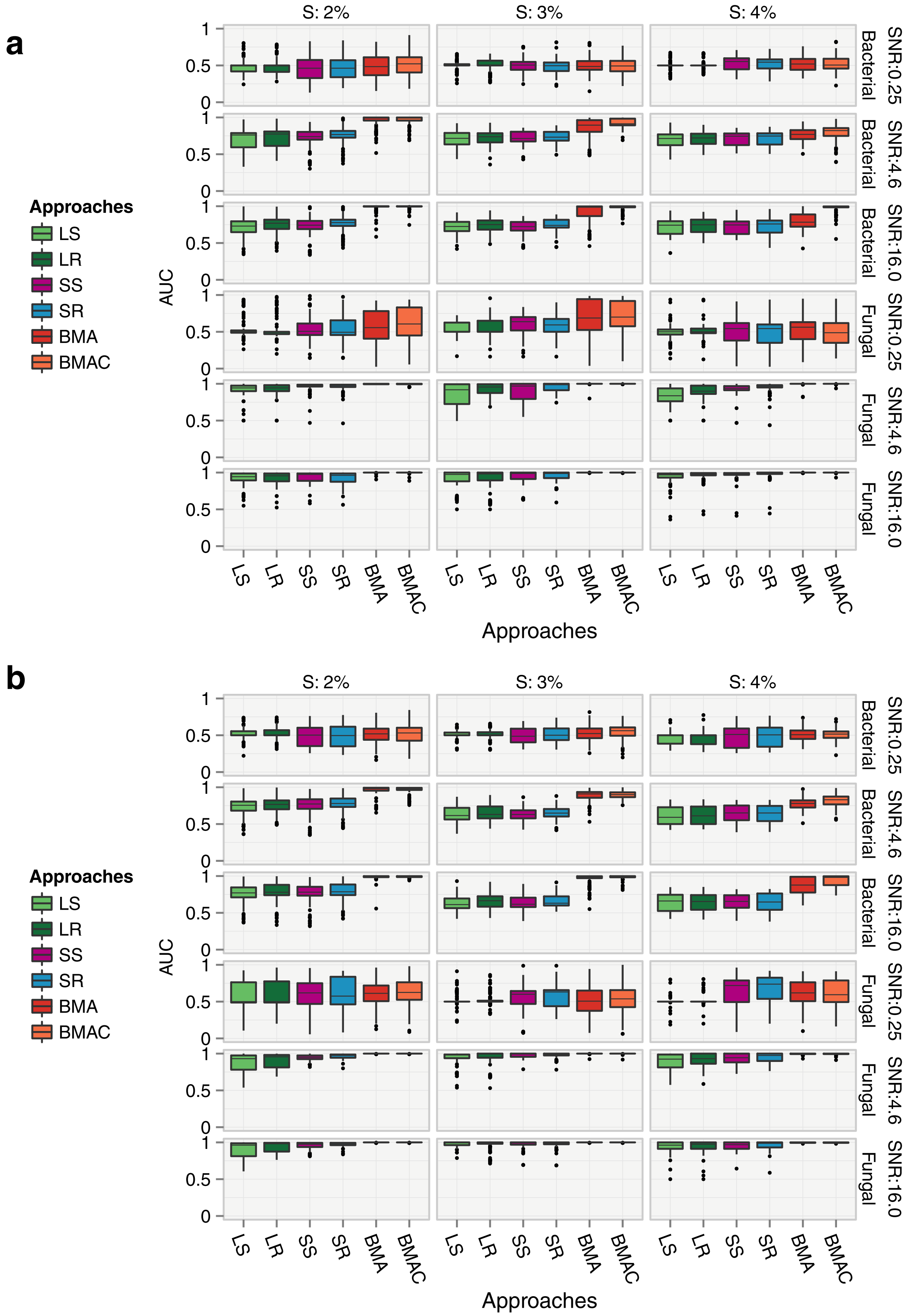


Figure 3 Variation in area under the ROC curve (AUC) across 130 simulations are shown for approaches that do not perform variable selection. Points beyond the end of the whiskers denote outliers. An AUC of 1.0 is ideal. *S* denotes the sparsity setting or the proportion of influential variables in the data. **a**. $\beta = \pm 1$ and **b**. $\beta \in \mathcal{U}$.

The lowest SNRs resulted in the lowest AUCs and the highest variability. Since SNRs of < 0.25 are not representative of our experimental datasets and in addition, yielded uniformly poor performance across all ensembles, we report results only for medium and high SNR situations in the evaluations that follow.

F-score

Higher values of the F-score indicate a good balance between precision and recall. Figure 4 shows the variation in F-scores across 130 simulation runs for all approaches while Additional file 1 shows the medians. For methods that did not perform variable selection, we applied the inclusion probability threshold from the LD algorithm to select the top influential variables.

BMAC outperformed all the other approaches across all settings. BMA was a close second, however, its performance deteriorated at lower sparsity settings. As was pointed out by one of our referees, this finding shows the benefits of cross-validating the *expected model size* within BMAC, enabling it to adapt better to the sparsity in the data. The PR algorithm performed as well as BMAC in the low-sparsity and medium SNR settings within high dimensions and in the presence of large regression coefficients ($\beta = \pm 1$). This would suggest that resampling conferred a distinct advantage over subsampling, enabling PR to perform well despite its small model space.

In lower dimensions, BMA approaches, with the largest ensembles, outperformed all other approaches across all SNR settings. SS and SR performed similar to BMA in a number of low-dimensional fungal settings but the variability in their performance was substantially higher.

Spearman's rank correlation

Figure 5 shows the variable ranking performance measured using Spearman's rank correlation between the estimated and true variable ranks for each of the approaches for $\beta \in \mathcal{U}$. The corresponding median correlations are shown in Additional file 1. Overall, the rank correlation decreased with an increase in dimensionality. BMA approaches showed higher correlation and lower variability relative to the other approaches in high dimensions. However, all approaches performed similarly in low dimensions.

Additional simulations

We performed two additional simulations to address points raised by our referees.

Performance without log-transformation

In this simulation, we examined the performance of all ensembles on relative abundance data that were not log-transformed. Data without log-transformation showed similar trends in performance as log-transformed data (Additional file 2). However, the number of outliers increased across all metrics in the data without log-transformation. This suggests that log-transformation of relative abundances decreases variability in the performance of ensembles.

Weighted average of inclusion probabilities

We also performed simulations to compare SS and SR with their weighted counterparts, SSW and SRW, computed using Equation 9. Overall, the performance of the weighted approach was not consistently different from the unweighted approach (Additional file 3). However, there were minor increases in the median performance of the weighted approach in high-SNR settings.

Application to mouse microbiome data

We now illustrate the application of the ensembles (Table 2) to identify influential variables in our experimental mouse microbiome data collected from the terminal ileum. Briefly, the gastrointestinal tract is an ecological niche for both bacterial and fungal flora. By depleting the competing bacterial microbiota, antibiotics create vacancies in the niche that co-existing fungi can repopulate. However, a high level of colonization by a single fungal species such as *C. albicans* is an unfavorable clinical outcome [48], especially in immunosuppressed and dysbiotic patient populations [49-51]. We therefore designed our mouse microbiome study to examine the factors influencing *C. albicans* colonization of the gastrointestinal tract following initiation of antibiotic treatment.

The experimental design of our mouse microbiome study is summarized in Figure 1 and Table 1. Following exposure to antibiotics and *C. albicans*, we profiled the microbial communities in the mouse gut by amplifying and sequencing the taxonomically discriminant bacterial 16S rDNA [52] from the V3-V5 variable region and the fungal ITS rDNA [53]. Using the YAP bioinformatic workflow [54], we obtained taxonomic assignments that reach genus-level resolution for both bacteria and fungi. Assignments unclassified at the genus-level were annotated with the prefix *UC*. The median number of reads per sample was 3500 for bacterial 16S and 2000 for fungal ITS regions. We built two linear regression models to assess the association of bacteria and fungi with the level of *C. albicans* colonization. We estimated the bacterial models from 30 samples and the fungal models from 36 samples. Both the bacterial and fungal models were specified as follows

$$Y = \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{X}_a\boldsymbol{\beta}_a + \boldsymbol{\epsilon}, \tag{21}$$

where Y is the colonization level measured in $log\left(\frac{\text{Colony forming units (CFU)}}{\text{grams of fecal matter}}\right)$, $\boldsymbol{X}_g$ is a vector of log relative abundances of sequences assigned to bacterial or

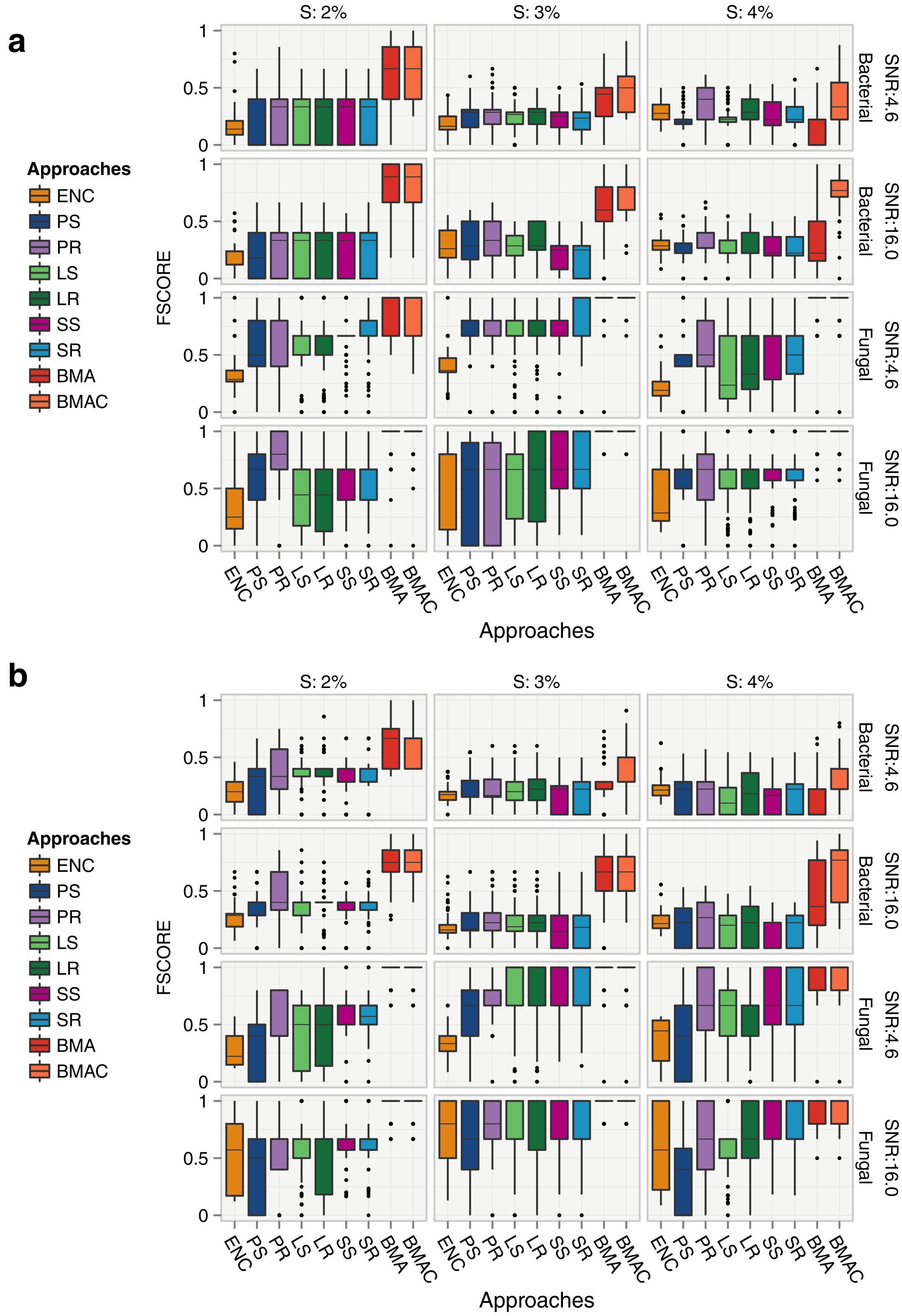


Figure 4 Variation in F-score across 130 simulations. Points beyond the end of the whiskers denote outliers. An F-score of 1.0 is ideal. The LD algorithm was used to select variables for approaches that do not perform variable selection. SNR=0.25 is not shown. *S* denotes the sparsity setting or the proportion of influential variables in the data. ***a***. $\beta = \pm 1$ and ***b***. $\beta \in \mathcal{U}$.

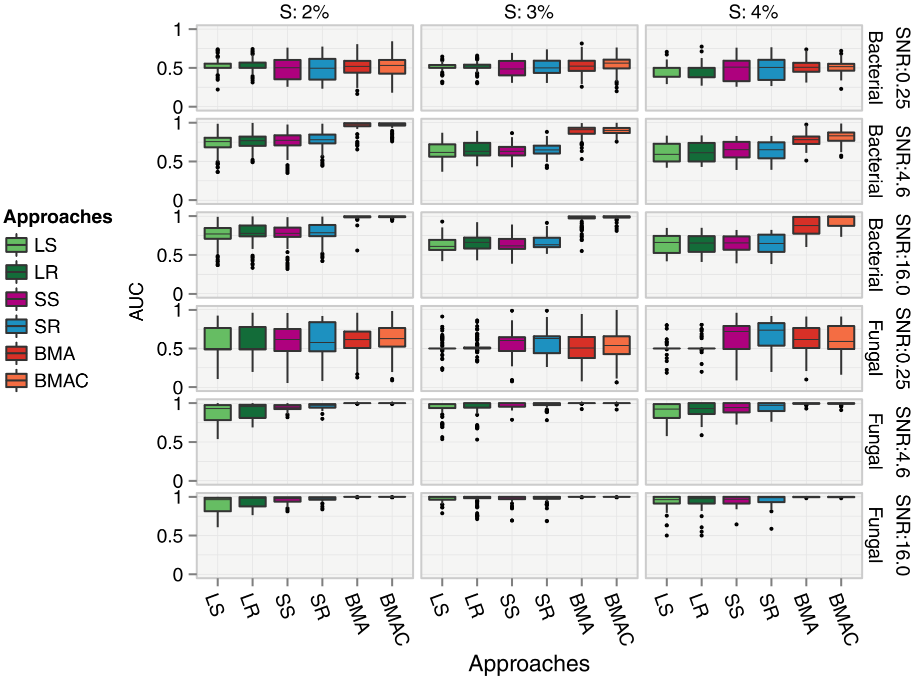


Figure 5 Variation in Spearman's rank correlation ($\beta \in \mathcal{U}$) across 130 simulations are shown for approaches that do not perform variable selection. Points beyond the end of the whiskers denote outliers. A correlation of 1.0 is ideal. Higher correlation indicates that the approach is able to capture the original variable rank more accurately. SNR=0.25 is not shown. *S* denotes the sparsity setting or the proportion of influential variables in the data.

fungal genera, $\boldsymbol{X_c}$ is a vector of log normalized mRNA expression levels of cytokines, $\boldsymbol{X_a}$ is a vector of 5 binary variables indicating antibiotic treatment (vancomycin or PSG: penicillin, streptomycin and gentamicin) for a span of 7 or 21 days and exposure to *C. albicans* on the 7th day, and $\boldsymbol{\epsilon}$ is the residual error in the model.

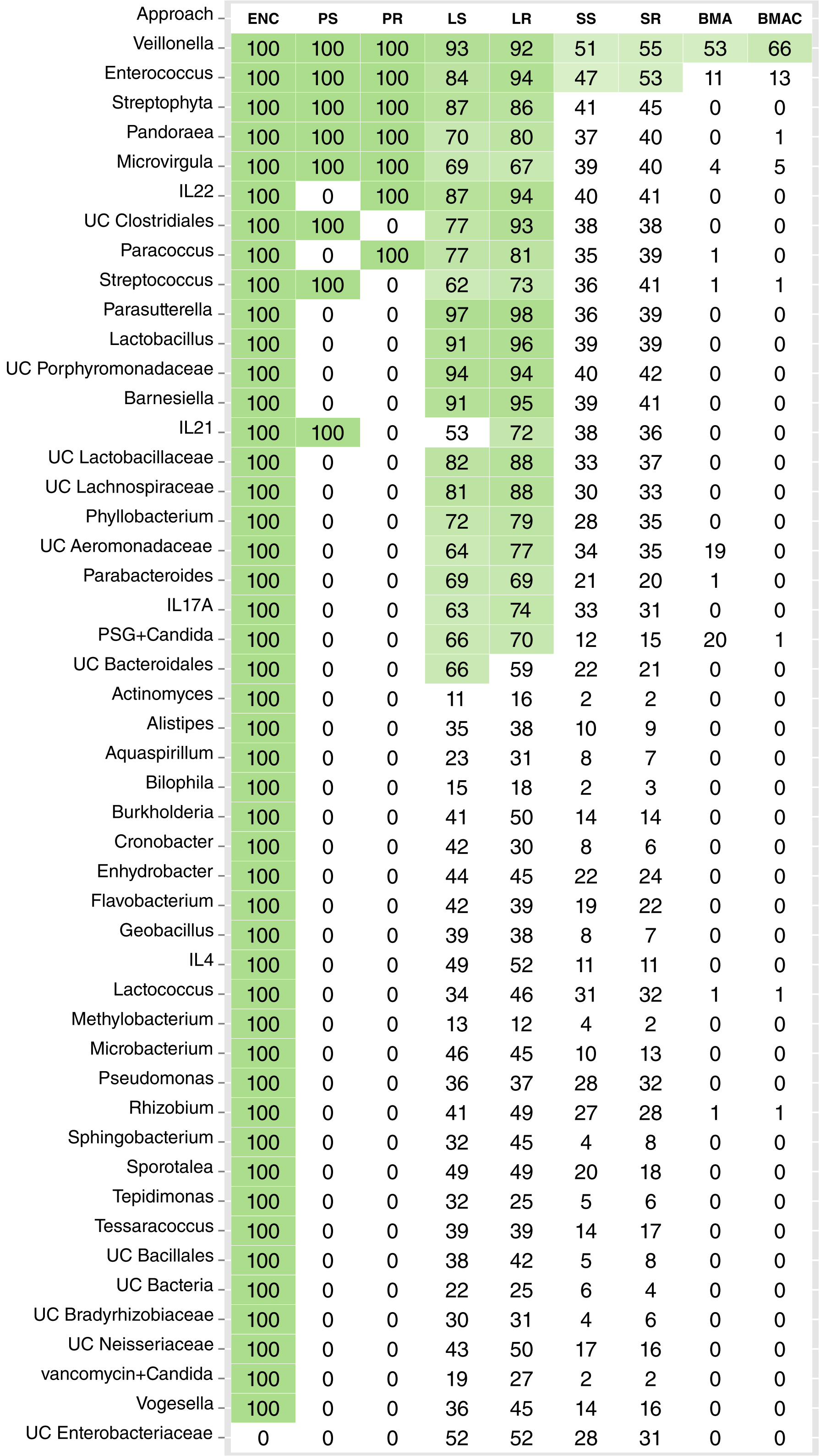
Since the microbial proportions sum to one, the proportion of any given genus is known given the proportions of all the other genera. To avoid this redundancy, we could exclude the proportion of any one reference genus. However, doing so would mean we are unable to obtain inclusion probability for this genus. We thus selected two reference genera with the highest abundances across all samples and repeated the model building procedure twice, each time excluding one of the two reference genera from $\boldsymbol{X_g}$. We estimated the final set of selected variables and inclusion probabilities for all genera by averaging across these two models.

We applied each of the approaches (Table 2) from the simulation study to this data to discover the microbiome-immune-antibiotic signatures associated with the level of *C. albicans* colonization. The inclusion probability for any variable reflects its importance to the response. For methods that do not perform variable selection, we used the LD algorithm to derive an inclusion probability threshold for identifying the top influential variables. Since ENC, PS and PR directly provide us with lists of influential variables and do not assign inclusion probabilities, we annotated the inclusion probabilities for these influential variables as 100%.

Mouse microbiome findings

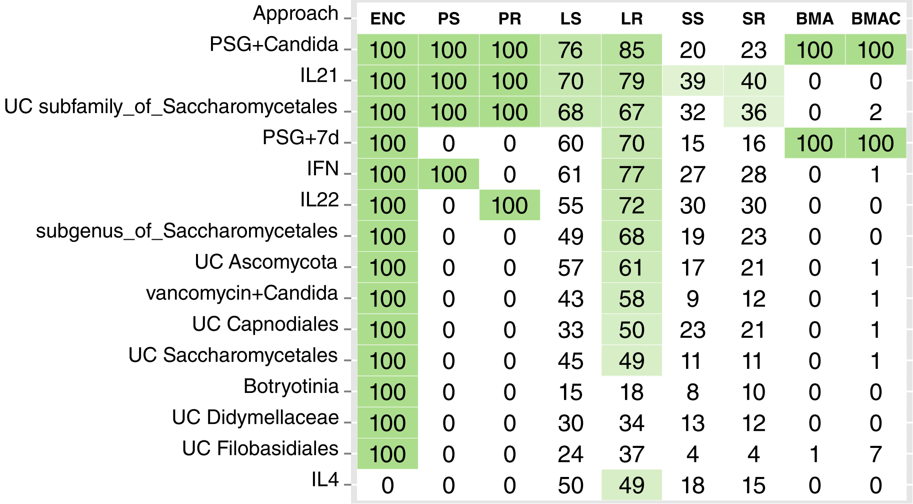
Figure 6 shows the bacteria and cytokines identified as influential by the approaches along with their inclusion probabilities (in %). ENC showed the least sparsity and selected the maximum number of bacteria and host immune cytokines. Although *Veillonella* was identified as an influential variable in all approaches, only BMA, BMAC, SS and SR assigned it the highest inclusion probability. However, unlike SS and SR, the difference in the inclusion probabilities of *Veillonella* and the other influential variables was much larger in the BMA models.

Figure 7 shows results from the fungal model. Treatment with PSG for a span of (a) 7 days, and (b) 21 days along with exposure to *C. albicans* were both assigned very high inclusion probabilities by BMA and BMAC. Unlike the bacterial model, the highest-ranking variables in SS and SR were different from those in BMA approaches. PR and PS shared common variables with both SS and SR and the BMA approaches. The other single λ approaches (LR, LS) showed some similarities with SS and SR, however, their inclusion probability



Approach	ENC	PS	PR	LS	LR	SS	SR	BMA	BMAC
Veillonella	100	100	100	93	92	51	55	53	66
Enterococcus	100	100	100	84	94	47	53	11	13
Streptophyta	100	100	100	87	86	41	45	0	0
Pandoraea	100	100	100	70	80	37	40	0	1
Microvirgula	100	100	100	69	67	39	40	4	5
IL22	100	0	100	87	94	40	41	0	0
UC Clostridiales	100	100	0	77	93	38	38	0	0
Paracoccus	100	0	100	77	81	35	39	1	0
Streptococcus	100	100	0	62	73	36	41	1	1
Parasutterella	100	0	0	97	98	36	39	0	0
Lactobacillus	100	0	0	91	96	39	39	0	0
UC Porphyromonadaceae	100	0	0	94	94	40	42	0	0
Barnesiella	100	0	0	91	95	39	41	0	0
IL21	100	100	0	53	72	38	36	0	0
UC Lactobacillaceae	100	0	0	82	88	33	37	0	0
UC Lachnospiraceae	100	0	0	81	88	30	33	0	0
Phyllobacterium	100	0	0	72	79	28	35	0	0
UC Aeromonadaceae	100	0	0	64	77	34	35	19	0
Parabacteroides	100	0	0	69	69	21	20	1	0
IL17A	100	0	0	63	74	33	31	0	0
PSG+Candida	100	0	0	66	70	12	15	20	1
UC Bacteroidales	100	0	0	66	59	22	21	0	0
Actinomyces	100	0	0	11	16	2	2	0	0
Alistipes	100	0	0	35	38	10	9	0	0
Aquaspirillum	100	0	0	23	31	8	7	0	0
Bilophila	100	0	0	15	18	2	3	0	0
Burkholderia	100	0	0	41	50	14	14	0	0
Cronobacter	100	0	0	42	30	8	6	0	0
Enhydrobacter	100	0	0	44	45	22	24	0	0
Flavobacterium	100	0	0	42	39	19	22	0	0
Geobacillus	100	0	0	39	38	8	7	0	0
IL4	100	0	0	49	52	11	11	0	0
Lactococcus	100	0	0	34	46	31	32	1	1
Methylobacterium	100	0	0	13	12	4	2	0	0
Microbacterium	100	0	0	46	45	10	13	0	0
Pseudomonas	100	0	0	36	37	28	32	0	0
Rhizobium	100	0	0	41	49	27	28	1	1
Sphingobacterium	100	0	0	32	45	4	8	0	0
Sporotalea	100	0	0	49	49	20	18	0	0
Tepidimonas	100	0	0	32	25	5	6	0	0
Tessaracoccus	100	0	0	39	39	14	17	0	0
UC Bacillales	100	0	0	38	42	5	8	0	0
UC Bacteria	100	0	0	22	25	6	4	0	0
UC Bradyrhizobiaceae	100	0	0	30	31	4	6	0	0
UC Neisseriaceae	100	0	0	43	50	17	16	0	0
vancomycin+Candida	100	0	0	19	27	2	2	0	0
Vogesella	100	0	0	36	45	14	16	0	0
UC Enterobacteriaceae	0	0	0	52	52	28	31	0	0

Figure 6 ***Candida*** **colonizatin: Influential variables in the bacterial model.** Variables selected in each of the modeling approaches are highlighted in green. The LD algorithm was utilized for selecting variables for approaches that do not perform variable selection. Values in the cells are inclusion probabilities expressed in %. ENC, PS and PR do not generate inclusion probabilities. In these models, an inclusion probability of 100% indicates that the variable was selected as influential.



Approach	ENC	PS	PR	LS	LR	SS	SR	BMA	BMAC
PSG+Candida	100	100	100	76	85	20	23	100	100
IL21	100	100	100	70	79	39	40	0	0
UC subfamily_of_Saccharomycetales	100	100	100	68	67	32	36	0	2
PSG+7d	100	0	0	60	70	15	16	100	100
IFN	100	100	0	61	77	27	28	0	1
IL22	100	0	100	55	72	30	30	0	0
subgenus_of_Saccharomycetales	100	0	0	49	68	19	23	0	0
UC Ascomycota	100	0	0	57	61	17	21	0	1
vancomycin+Candida	100	0	0	43	58	9	12	0	1
UC Capnodiales	100	0	0	33	50	23	21	0	1
UC Saccharomycetales	100	0	0	45	49	11	11	0	1
Botryotinia	100	0	0	15	18	8	10	0	0
UC Didymellaceae	100	0	0	30	34	13	12	0	0
UC Filobasidiales	100	0	0	24	37	4	4	1	7
IL4	0	0	0	50	49	18	15	0	0

Figure 7 ***Candida*** **colonization: Influential variables in the fungal model.** Variables selected in each of the modeling approaches are highlighted in green. The LD algorithm was utilized for selecting variables for approaches that do not perform variable selection. Values in the cells are inclusion probabilities expressed in %. ENC, PS and PR do not generate inclusion probabilities. In these models, an inclusion probability of 100% indicates that the variable was selected as influential.

assignments were substantially different. ENC showed a lack of sparsity analogous to the bacterial models.

Viewed together, findings from the bacterial and fungal models indicate that the type of antibiotic, span of exposure to the antibiotic and specific antibiotic-induced changes in the bacterial microbiota substantially influenced the levels of *C. albicans* colonization. The concurrent effects of the host immune cytokines and the co-existing fungal microbiota on the level of *C. albicans* colonization were weaker. These findings are promising since they suggest that tailoring antibiotic regimens as well as concomitant modulation of the microbiota during antibiotic administration could be employed to reduce opportunistic *C. albicans* colonization.

Comparison of approaches

To further examine similarities and differences in the workings of each of the modeling approaches, we plotted the inclusion probability versus rank for all variables across the ensembles (Figure 8). In both high and low-dimensional settings, the approaches separated out into three distinct groups. In the first group that included BMA and BMAC, inclusion probabilities showed the steepest decay. In the second, comprising the SS and SR, the decay was more gradual. The single λ approaches constituted the third group. Here the rate of decay was similar to SS and SR, however the dynamic range of inclusion probabilities was larger. The sharper decay in the BMA methods indicates a more discriminative ranking that allows a distinct separation of the influential variables from the non-influential ones.

On clustering the approaches based on Euclidean distance between the inclusion probability vectors, the single λ approaches were further away from PS, PR, SS, SR and the BMA approaches, indicating their divergent performance. The proximity of SS and SR to BMA approaches suggests that SS and SR, computed over a much larger number of subsamples or resamples, could potentially approach the performance profile of BMA.

Conclusions

Microbiome studies constitute a recently developed area of research that incorporate NGS to profile communities of diverse microbes residing in a variety of ecosystems. As a result, microbiome data shares many of the characteristics of NGS data that make it challenging to analyze. These challenges arise from high-dimensionality, limited sample sizes, multicollinearity within covariables, modest SNRs and a very large space of models. Ensemble modeling is able to provide a comprehensive solution to these analytical challenges. However, except for non-linear ensembles such as random forests, ensemble modeling has not been adopted widely in microbiome research.

Our goal in this paper was to familiarize investigators in the field with the characteristics of both frequentist and Bayesian ensemble-based regression approaches and present an evaluation of these approaches on realistic microbiome data. We also introduce a simulation framework that can be built from any experimental microbiome dataset to closely mimic the characteristics of real data. We demonstrate the usefulness of this framework in evaluating analytical approaches using a suite of metrics

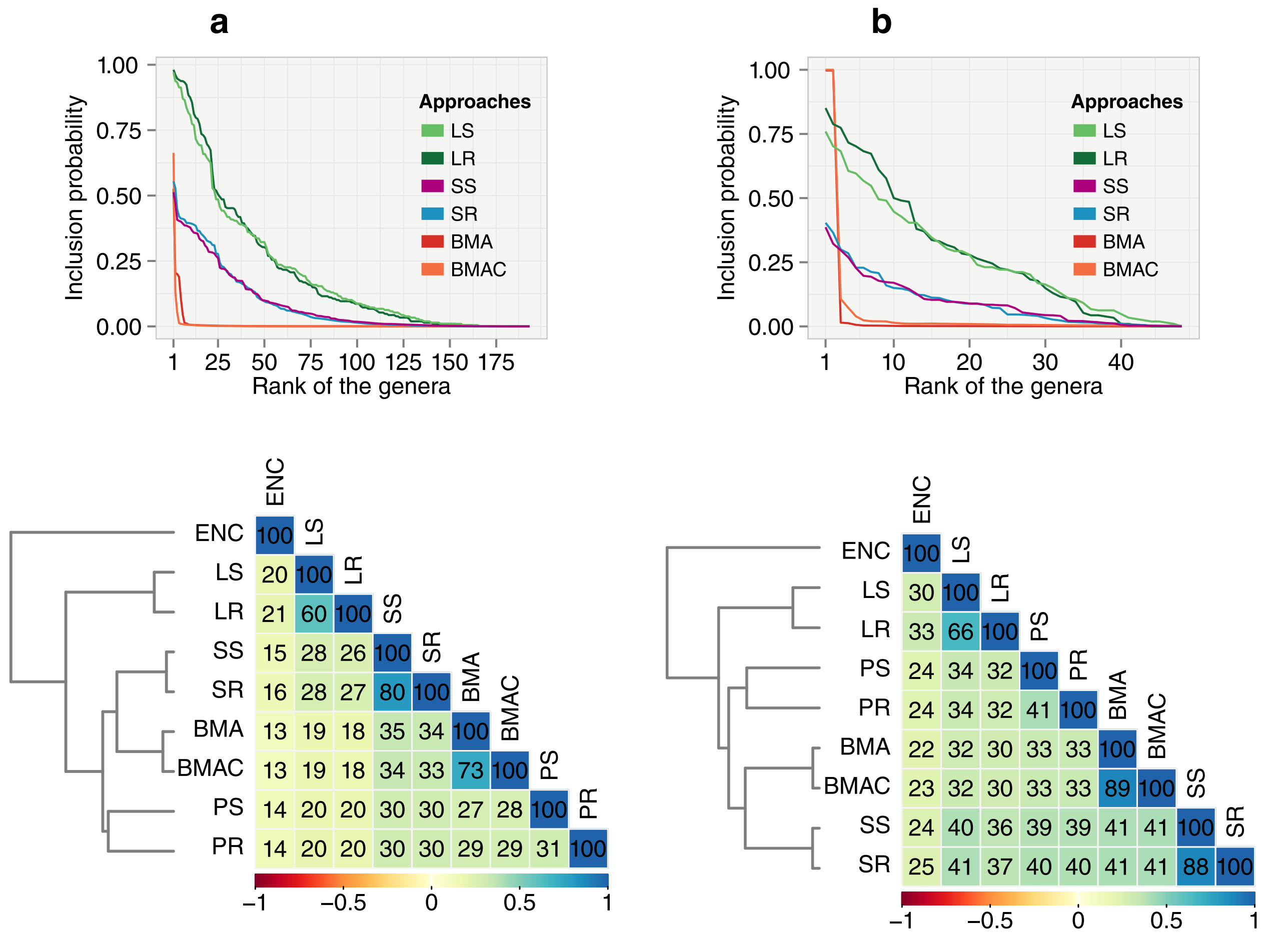


Figure 8 Comparison of inclusion probability profiles across all approaches. The top panels show the sorted inclusion probabilities plotted against the rank of the variables in the model. The bottom panels show the pairwise similarities between the inclusion probability profiles of all the approaches. The similarity between any pair of approaches *i* and *j* is given as $S_{ij} = \frac{1}{1+D_{ij}}$, where D_{ij} is the Euclidean distance between the inclusion probability vectors of approaches *i* and *j*. We generated the dendrogram using agglomerative clustering with the centroid linkage method. We then ordered the correlogram using the order of models generated in the dendrogram. **a**. Bacterial models, **b**. Fungal models.

that assess various facets of modeling performance. Unlike other simulation strategies [4,16] that sample covariates from parametric distributions, we directly utilize the covariance structure, SNR and the dimensions inherent in our experimental mouse data for our simulation. Even though parametric simulation approaches enable generation of arbitrary number of samples, we expect the findings from our simulation-based evaluation to translate better to experimental datasets. Our implementation, in the R statistical language, is freely available from *GitHub* [30], making it convenient for researchers in the microbiome research community to adopt and apply our evaluation methodology to their own datasets without the need for expensive ground truth annotations.

We found that the Bayesian approaches had several favorable attributes that distinguished them from the other ensembles we evaluated. In realistic settings with medium SNRs and uniformly distributed coefficients for influential variables ($\beta \in \mathcal{U}$), they performed consistently better or at par with frequentist approaches depending on the number of influential variables. Furthermore, they showed substantially less variation across replications. Selecting the optimal *expected model size* via cross-validation (BMAC) improved performance in more challenging high-dimensional settings with a large number of influential variables. PR, the resampling variant of the original stability selection algorithm [22] performed as well as BMAC in these scenarios, highlighting its ability to perform effective variable selection when there are several variables with large regression coefficients. On the whole, the resampling variants had small but consistent performance gains over their subsampling counterparts, suggesting that resampling was able to create an improved model space relative to subsampling. Even though SS, SR, PS and PR, the frequentist ensembles, performed reasonably well at lower dimensions, they exhibited substantially more variability in their performance. Our findings thus enabled us to pick BMAC for our experimental setting with modest sample sizes and high-dimensionality in measurements.

We note that BMAC uses the data for both estimating a hyperparameter of the prior distribution and for computing the posterior distribution. It is therefore not a full Bayesian approach. An ideal alternative would be to employ a full Bayesian model that includes a hyperprior for the prior inclusion probabilities [55]. While our work has systematically evaluated a powerful set of sparse linear regression modeling methods that average across large model spaces in the context of continuous response variables, a future direction would be to extend these evaluations for analyzing categorical responses and time-to-event outcomes that are prevalent in clinical settings. These models include logistic, multinomial, and survival regression.

Our analysis could have been performed at the higher taxonomic levels of phylum, class, order and family. However, we selected the genus-level primarily because it provided the closest taxonomic resolution to our response (*C. albicans* colonization). A limitation of our approach, which is shared by other analytical approaches such as [4,16,17], is the assumption that taxonomic assignment for all organisms at all phylogenetic levels is equally accurate. As a result, covariables are treated alike regardless of classification accuracy. However in practice, classification error could vary across phylogenetic branches as well as taxonomic levels and depends on factors such as the quality of annotated taxonomic databases as well as the other parameters of underlying bioinformatic algorithms [56]. A crucial avenue of future research would be to integrate information from models constructed from taxonomic assignments at varying classification accuracies.

Another area that we have not explored in this work is the inclusion of interactions and variable sub-structures within the high-dimensional regression framework. In the microbiome setting, this would involve several higher-order interaction terms among taxa. Many such symbiotic and antagonistic relationships are known to exist among taxa and cytokines but are hard to assess in analytical settings. Including these interactions is computationally challenging because they would increase the model space exponentially. Recent developments in statistical algorithms that efficiently explore this additional complex model space [57,58] hold promise for making this problem more tractable at the scale of microbiome data. In the context of biological interpretations, it could be useful to integrate variable clustering with ensemble-based regression in a framework similar to the one proposed in [59] to obtain a greater understanding of the most relevant dynamics and relationships in the community.

In conclusion, as microbiome studies evolve towards translational settings, analysts are likely to face challenges in selecting appropriate modeling strategies that yield consistent and stable performance with low variability [60]. Therefore, we expect that our research will provide both insights for choosing among ensemble methods and an evaluation framework critical for making an objective selection.

Additional files

Additional file 1: Median values of performance metrics (median_scores.pdf). Figures showing median values of performance metrics presented as boxplots in the main paper.

Additional file 2: Performance without log-transformation (no_log_transformation.pdf). Figures showing performance of all the approaches on data without log-transformations.

Additional file 3: Performance with weighted inclusion probabilities (weighted_ip.pdf). Figures showing performance of SS, SR, SSW and SRW approaches.

Abbreviations

AUC: Area under the ROC curve; BMA: Bayesian model averaging; BMAC: Bayesian model averaging with cross-validation for expected model size; *Bolasso*: Bootstrap-enhanced LASSO; CFU: Colony forming units; CV: Cross-validation; EMS: Expected model size; ENC: Elastic net with cross-validation; FP: False positives; FPR: False positive rate; GAPDH: Glyceraldehyde-3-Phosphate Dehydrogenase; GI: Gastrointestinal; ITS: Internal transcribed spacer; LARS: Least angle regression; LASSO: Least absolute shrinkage and selection operator; LD: Lagged differences; LR: Stability selection on resamples using λ_{ENC}; LS: Stability selection on subsamples using λ_{ENC}; MB: Meinshausen & Bühlmann; MCMC: Markov chain Monte Carlo; MSE: Mean squared error; NGS: Next-generation sequencing; NIH: National Institutes of Health; PR: Stability selection on resamples minimizing expected number of false positives; PS: Stability selection on subsamples minimizing expected number of false positives; PSG: penicillin, streptomycin, gentamicin; ROC: Receiver operating characteristic; RSS: Residual sum of squares; SNR: Signal to noise ratio; SR: Stability selection on resamples using unweighted average inclusion probabilities across $\lambda \in \Lambda$; SRW: Stability selection on resamples using weighted average inclusion probabilities across $\lambda \in \Lambda$; SS: Stability selection on subsamples using unweighted average inclusion probabilities across $\lambda \in \Lambda$; SSW: Stability selection on subsamples using weighted average inclusion probabilities across $\lambda \in \Lambda$; TP: True positives; TPR: True positive rate; UC: Unclassified.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JS designed and implemented the simulations, performed the evaluation experiments and wrote the manuscript. SS designed and implemented the YAP bioinformatics workflow for taxonomic assignment and provided crucial insights into biases and structure of microbiome data. NVS and HL conducted the mouse experiments. SM conducted the microbiome sequencing experiments. SGF, WCN, NVS, LL and SM designed the mouse microbiome study. All the authors read and approved the manuscript.

Acknowledgements

We gratefully acknowledge our two anonymous referees whose insightful comments substantially improved our manuscript. We thank the authors of *glmnet, BoomSpikeSlab, quadrupen, scales, ggplot2, plyr, reshape2, data.table, Hmisc, doMC, corrplot, RColorBrewer, RStudio* and the R Core Development Team for their invaluable contributions to statistical computing. We thank

Karla Stucker and Andrey Tovchigrechko for their constructive discussions and Reed Shabman for contributing *"Reed's mouse"* to our manuscript. This research was supported by NIH grants R01AI054928 and R01DE017088 and the Bill and Melinda Gates foundation grant, OPP1017579.

Author details

[1] J. Craig Venter Institute, 9704, Medical Center Drive, Rockville, 20850 Maryland, US. [2] Los Angeles Biomedical Research Institute at Harbor, UCLA Medical Center, 1124 West Carson Street, Torrance, 90509 California, US. [3] David Geffen School of Medicine, University of California at Los Angeles, 90095 California, US.

Received: 8 July 2014 Accepted: 15 January 2015
Published online: 01 February 2015

References

1. The NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH human microbiome project. Genome Res. 2009;19(12):2317–23. doi:10.1101/gr.096651.109.
2. Hood L. Tackling the microbiome. Science. 2012;336(6086):1209. doi:10.1126/science.1225475.
3. Gilbert JA, Bailey M, Field D, Fierer N, Fuhrman JA, Hu B, et al. The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2011. Standards Genomic Sci. 2011;5(2):243–7. doi:10.4056/sigs.2134923.
4. Chen J, Li H. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. Ann Appl Stat. 2013;7(1): 418–42. doi:10.1214/12-aoas592.
5. Xia F, Chen J, Fung WK, Li H. A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics. 2013;69(4):1053–63. doi:10.1111/biom.12079.
6. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009;5(4):1000352. doi:10.1371/journal.pcbi.1000352.
7. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2. doi:10.1038/nmeth.2658.
8. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett W, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):60. doi:10.1186/gb-2011-12-6-r60.
9. Taur Y, Xavier JB, Lipuma L, Ubeda C, Goldberg J, Gobourne A, et al. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. Clin Infectious Diseases. 2012;55(7):905–14. doi:10.1093/cid/cis580.
10. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12): 8228–35. doi:10.1128/aem.71.12.8228-8235.2005.
11. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106–13. doi:10.1093/bioinformatics/bts342.
12. Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One. 2014;9(2):87830. doi:10.1371/journal.pone.0087830.
13. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343–59. doi:10.1111/j.1574-6976.2010.00251.x.
14. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1(11):. doi:10.1186/2049-2618-1-11.
15. Waldron L, Pintilie M, Tsao M-SS, Shepherd FA, Huttenhower C, Jurisica I. Optimized application of penalized regression methods to diverse genomic data. Bioinformatics. 2011;27(24):3399–406. doi:10.1093/bioinformatics/btr591.
16. Garcia TP, Müller S, Carroll RJ, Walzem RL. Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. Bioinformatics. 2014;30(6):831–7. doi:10.1093/bioinformatics/btt608.
17. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. Biometrika. 2014;101(4):785–797. doi:10.1093/biomet/asu031.
18. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2009;367(1906):4237–53. doi:10.1098/rsta.2009.0159.
19. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R, 1st edn. New York, NY, USA: Springer Series in Statistics, Springer; 2013. doi:10.1007/978-1-4614-7138-7.
20. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231. doi:10.1214/ss/1009213726.
21. Yu B. Stability. Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability. 2013;19(4):1484–1500. doi:10.3150/13-bejsp14.
22. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B, Stat Methodology. 2010;72(4):417–73. doi:10.1111/j.1467-9868.2010.00740.x.
23. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Stat Sci. 1999;14(4):382–417. doi:10.1214/ss/1009212814.
24. George EI, McCulloch RE. Approaches for bayesian variable selection. Statistica Sinica. 1997;7:339–73.
25. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
26. Mosteller F, Tukey JW. Data Analysis and Regression - A Second Course in Statistics, Student. edn. Reading, MA: Addison-Wesley Publishing Company; 1977.
27. Bühlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications, 1st edn. Springer, Heidelberg, Germany: Springer Series in Statistics; 2011. doi:10.1007/978-3-642-20192-9.
28. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B, Stat Methodology. 2005;67(2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.
29. R Development Core Team. R: A Language and Environment for Statistical Computing. 2014. http://www.R-project.org/. Accessed on 07 February, 2015.
30. Shankar J. *regeval*: A systematic evaluation of high-dimensional, ensemble based regression for exploring large model spaces in microbiome analyses. GitHub. 2014. http://github.com/openpencil/regeval. Accessed on 07 February, 2015.
31. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, NY, USA: Springer Series in Statistics; 2013. doi:10.1007/978-0-387-84858-7.
32. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software. 2010;33(1):1–22.
33. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B, Stat Methodology. 1996;58(1):267–88. doi:10.2307/2346178.
34. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32(2):407–99. doi:10.1214/009053604000000067.
35. Grandvalet Y, Chiquet J, Ambroise C. Sparsity by Worst-Case Quadratic Penalties. arXiv preprint. 2012. http://arxiv.org/abs/1210.2077 Accessed on 07 February, 2015. arXiv:1210.2077 Code: http://r-forge.r-project.org/projects/quadrupen/ Accessed on 07 February, 2015.
36. Mason DM, Newton MA. A rank statistics approach to the consistency of a general bootstrap. Ann Stat. 1992;20(3):1611–24. doi:10.1214/aos/1176348787.
37. Praestgaard J, Wellner JA. Exchangeably weighted bootstraps of the general empirical process. Ann Probability. 1993;21(4):2053–86. doi:10.1214/aop/1176989011.
38. Freedman D. A remark on the difference between sampling with and without replacement. J Am Stat Assoc. 1977;72(359):681–1. doi:10.1080/01621459.1977.10480637.
39. Bühlmann P, Yu B. Analyzing bagging. Ann Stat. 2002;30(4):927–61. doi:10.1214/aos/1031689014.
40. Friedman JH, Hall P. On bagging and nonlinear estimation. J Stat Planning Inference. 2007;137(3):669–83. doi:10.1016/j.jspi.2006.06.002.
41. Dümbgen L, Samworth RJ, Schuhmacher D. Stochastic search for semiparametric linear regression models. In: Banerjee M, Bunea F, Huang J, Koltchinskii V, Maathuis MH, (eds.) From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner. Beachwood, Ohio, USA: Institute of Mathematical Statistics; 2013. p. 78–90. doi:10.1214/12-IMSCOLL907.

42. Bach FR. Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08. New York, New York, USA: ACM Press; 2008. p. 33–40. doi:10.1145/1390156.1390161.
43. Scott SL. *BoomSpikeSlab*: MCMC for spike and slab regression. R package version 0.5.2 CRAN. 2014. http://CRAN.R-project.org/package=BoomSpikeSlab Accessed on 07 February, 2015.
44. Scott SL, Varian HR. Predicting the present with bayesian structural time series. Int J Math Modell Numer Optimisation. 2014;5(1/2):4. doi:10.1504/ijmmno.2014.059942.
45. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using Bayesian structural time-series models. Ann Appl Stat. 2014. In press. http://www.e-publications.org/ims/submission/AOAS/user/submissionFile/17112?confirm=ef7e4bee. Accessed on 07 February, 2015.
46. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g priors for bayesian variable selection. J Am Stat Assoc. 2008;103(481):410–23. doi:10.1198/016214507000001337.
47. Cortes C, Mohri M. AUC optimization vs. error rate minimization. In: Thrun S, Saul LK, Schölkopf B, Thrun S (eds.) Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT Press; 2004. p. 313–20.
48. Das I, Nightingale P, Patel M, Jumaa P. Epidemiology, clinical characteristics, and outcome of candidemia: experience in a tertiary referral center in the UK. Int J Infectious Dis. 2011;15(11):e759–63. doi:10.1016/j.ijid.2011.06.006.
49. Smeekens SP, van de Veerdonk FL, Kullberg BJ, Netea MG. Genetic susceptibility to candida infections. EMBO Mol Med. 2013;5(6):805–13. doi:10.1002/emmm.201201678.
50. Giri S, Kindo AJ. A review of candida species causing blood stream infection. Indian J Med Microbiol. 2012;30(3):270–8. doi:10.4103/0255-0857.99484.
51. Eggimann P, Bille J, Marchetti O. Diagnosis of invasive candidiasis in the ICU. Ann Intensive Care. 2011;1(1):37. doi:10.1186/2110-5820-1-37.
52. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen K. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infection. 2008;14(10):908–34. doi:10.1111/j.1469-0691.2008.02070.x.
53. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List: Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc Nat Acad Sci USA. 2012;109(16):6241–6. doi:10.1073/pnas.1117018109.
54. Szpakowski S. YAP: A Computationally Efficient Workflow for Taxonomic Analyses of Bacterial 16S and Fungal ITS Sequences. GitHub. 2013. http://github.com/shpakoo/YAP. Accessed on 07 February, 2015.
55. Scott JG, Berger JO. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. Ann Stat. 2010;38(5):2587–619. doi:10.1214/10-AOS792.
56. May A, Abeln S, Crielaard W, Heringa J, Brandt BW. Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. Bioinformatics. 2014;30(11):1530–8. doi:10.1093/bioinformatics/btu085.
57. Shah RD. Modelling interactions in high-dimensional data with Backtracking. arXiv preprint. 2013. http://arxiv.org/abs/1208.1174. Accessed on 07 February, 2015.
58. Shah RD, Meinshausen N. Random intersection trees. J Machine Learning Res. 2014;15(1):629–54.
59. Witten DM, Shojaie A, Zhang F. The cluster elastic net for High-Dimensional regression with unknown variable grouping. Technometrics. 2013;56(1):112–22. doi:10.1080/00401706.2013.810174.
60. Prosser JI. Replicate or lie: The need for replication. Environ Microbiol. 2010;12(7):1806–10. doi:10.1111/j.1462-2920.2010.02201.x.