Bioinformatics, 31(20), 2015, 3322–3329
doi: 10.1093/bioinformatics/btv364
Advance Access Publication Date: 16 June 2015
Original Paper

OXFORD

Systems biology

Investigating microbial co-occurrence patterns based on metagenomic compositional data

Yuguang Ban[1], Lingling An[2,3] and Hongmei Jiang[1,*]

[1]Department of Statistics, Northwestern University, Evanston, IL 60208, USA, [2]Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721, USA and [3]Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ 85721, USA

*To whom correspondence should be addressed.

Associate editor: Igor Jurisica

Received on March 31, 2015; revised on May 25, 2015; accepted on June 6, 2015

Abstract

Motivation: The high-throughput sequencing technologies have provided a powerful tool to study the microbial organisms living in various environments. Characterizing microbial interactions can give us insights into how they live and work together as a community. Metagonomic data are usually summarized in a compositional fashion due to varying sampling/sequencing depths from one sample to another. We study the co-occurrence patterns of microbial organisms using their relative abundance information. Analyzing compositional data using conventional correlation methods has been shown prone to bias that leads to artifactual correlations.

Results: We propose a novel method, regularized estimation of the basis covariance based on compositional data (REBACCA), to identify significant co-occurrence patterns by finding sparse solutions to a system with a deficient rank. To be specific, we construct the system using log ratios of count or proportion data and solve the system using the l_1-norm shrinkage method. Our comprehensive simulation studies show that REBACCA (i) achieves higher accuracy in general than the existing methods when a sparse condition is satisfied; (ii) controls the false positives at a prespecified level, while other methods fail in various cases and (iii) runs considerably faster than the existing comparable method. REBACCA is also applied to several real metagenomic datasets.

Availability and implementation: The R codes for the proposed method are available at http://faculty.wcas.northwestern.edu/~hji403/REBACCA.htm

Contact: hongmei@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microorganisms coexist with highly diverse patterns in different environments. Studying the microbial inter-organism relationships may provide us important insights to the underlying properties of the ecosystem. It has been shown that symbiosis of microbes is responsible for important metabolic processes such as pesticide degradation (Katsuyama *et al.*, 2009), biogeochemical cycling in seawater (Orphan *et al.*, 2001) and dental plague development (Hojo *et al.*, 2009). With recent development of high-throughput sequencing technologies, metagenomic studies based on uncultivated microbial samples have allowed us to survey the compositions of microbial communities. Although numerous microbes have been identified, our understandings of the relationships among microbes have been much less fruitful. One of the reasons is that we lack of tools to analyze the correlation structure in metagenomic data due to its high dimensions and complex distributions.

After annotating/binning of metagenomic sequencing data, the outputs are usually recorded as counts for downstream analysis. Because metagenomic samples usually cannot be collected at the same scale, after identifying the taxa or operational taxonomic units (OTUs) in a sample, the count of each OTU is usually then converted into relative abundance such as proportion or percentage by

dividing its count by the total. This implies that the metagenomic relative abundance data follow a compositional fashion (Aitchison, 1981). It is well known that applying conventional correlation coefficient methods on compositional data may lead to biased results (Lovell *et al*., 2010), although the compositional effect may be weaker on datasets with larger number of components (Friedman and Alm, 2012), such as the metagenomic compositional data. To take advantage of the existing statistical methods and also avoid producing spurious correlations, various types of data transformations have been applied in many compositional data studies. These transformations include the additive and centered log ratio transformation (Aitchison, 1986), isometric log ratio transformation (Egozcue *et al*., 2003). Chen and Li (2013) modeled the metagenomic data directly using Dirichlet-Multinomial distribution. Recently, Kurtz *et al*. (2015) studied the microbial ecological networks using the concept of conditional independence. These studies provide us tools to handle compositionality, however, using these tools to infer pair-wise correlations has been a challenge.

Some computational techniques have been developed to mitigate the compositional effect. To evaluate the significance of Pearson correlation coefficient calculated based on compositional data, Faust *et al*. (2012) used a permutation-renormalization bootstrap method (ReBoot) to mitigate the compositional bias. Using simulation studies, we find that while ReBoot shows some improvement over the conventional method, its performance is not consistent and depends on data structure. By assuming unobserved basis abundance, SparCC was proposed to infer the pair-wise correlations of basis abundance rather than their proportions (Friedman and Alm, 2012). Basis abundance is defined as a positive unconstrained quantity which forms the composition. Although basis abundance is conceptual, it does not involve any relative information as opposed to relative abundance, and thus analysis based on it is free of the compositional bias. SparCC estimates correlations based on basis abundance but is not efficient due to its high computational complexity. Therefore, these methods have their limitations and have not been investigated thoroughly against compositional data with various types of structures.

We propose a method, *r*egularized *e*stimation of the *ba*sis *c*ovariance based on *c*ompositional d*a*ta (REBACCA), which estimates the correlations between pairs of basis abundance using the log ratio transformation of metagenomic count or proportion data. Estimating the basis covariance structure from compositional data is equivalent to solving a linear system with a deficient rank. Difference between REBACCA and SparCC can be summarized as: (i) we construct a linear system that is exactly equivalent to the log ratio transformations and (ii) we use the popular l_1-norm shrinkage method to solve the system under a sparsity assumption. Our simulation studies show that REBACCA achieves higher accuracy in general and has better asymptotic performance with large sample size as compared with other methods. Its performance is more consistent on datasets of various structures. It is also computationally efficient and can be used to analyze large-scale metagenomic data.

2 Materials and Methods

The REBACCA scheme is designed to infer covariance structure of the basis abundance, given observed count or proportion data. To evaluate the method we perform simulation and real data analyses. The simulation process mimics the sampling procedure in obtaining real data from collecting to sequencing microbiome samples. The real metagenomic datasets are from Srinivas *et al* (2013), where microbial data are collected from three groups of mouse skin samples.

2.1 Regularized estimation of the basis covariance based on compositional data

REBACCA assumes that the basis abundances of the microbes in a microbiome population are unknown, and the observed data are either counts or proportions of the taxa or OTUs contained in a metagenomic sample. REBACCA mainly consists of two parts. It first constructs a linear system using log ratios between pairs of compositions, and then utilizes l_1-norm penalty to solve the system, which is rank-deficient due to the log ratio transformation.

2.1.1 Estimating log basis covariance structure

Suppose the unobserved basis abundances of D taxa $\mathbf{x} = (x_1, \ldots, x_D)$ are random variables and let $\mathbb{\Sigma} = (\sigma_{ij})_{D\times D}$ be the variance-covariance matrix of $\log(\mathbf{x})$. For any two taxa, i and j, we have

$$var\left[\log\left(x_i/x_j\right)\right] = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}. \quad (1)$$

As the basis abundances are unobserved we cannot estimate σ_{ij} directly. However, we can estimate $var\left[\log\left(x_i/x_j\right)\right]$ using the observed count or proportion data. To avoid undefined log ratios, zero values in the data are replaced by a small value equal to 1/10 of the minimum of non-zero values. While our goal is to estimate $\mathbb{\Sigma}$ in (1), it is generally impossible to find a unique solution without knowing the structure of $\mathbb{\Sigma}$. Aitchison studied the independent case where $\mathbb{\Sigma}$ is of a diagonal structure (Aitchison, 1981). Friedman and Alm (2012) introduced sparse assumption on $\mathbb{\Sigma}$ in SparCC, however they did not clearly specify the sparse structure of $\mathbb{\Sigma}$. Yet, SparCC estimates $\mathbb{\Sigma}$ by refining its solution recursively using a correlation threshold, which is computationally inefficient. Here, we develop a different framework to utilize the fast l_1-norm shrinkage method to estimate $\mathbb{\Sigma}$. We also discuss the sparse condition under which the estimation is accurate. We consider that the off-diagonal elements of $\mathbb{\Sigma}$ are unknown variables, and our goal is to identify and estimate the non-zero ones. We construct such a system as follows.

Summing up (1) on both sides, we have

$$\sum_{i\neq j} var\left[\log\left(x_i/x_j\right)\right] = \sum_{i\neq j}(\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}). \quad (2)$$

Let $s(\mathbf{x}) = \sum_{i\neq j} var\left[\log\left(x_i/x_j\right)\right]$. Define a series of vectors based on the ratios, $\mathbf{y}_d = (\log\left(\frac{x_1}{x_d}\right), \ldots, \log\left(\frac{x_{d-1}}{x_d}\right), \log\left(\frac{x_{d+1}}{x_d}\right), \ldots, \log\left(\frac{x_D}{x_d}\right))$ for $d = 1, \ldots, D$, and the corresponding variance-covariance matrix $\Omega_d(\mathbf{x}) = var(\mathbf{y}_d)$. It can be seen that $s(\mathbf{x}) = \sum_{d\in\{1,\ldots,D\}} \mathrm{trace}(\Omega_d(\mathbf{x}))$. Denote by $\mathbf{x}_{-t}$ the random variable $\mathbf{x}$ excluding the tth taxon. Then similarly we can define $\Omega_d(\mathbf{x}_{-t})$ for $d \neq t$ and $s(\mathbf{x}_{-t}) = \sum_{d\neq t} \mathrm{trace}(\Omega_d(\mathbf{x}_{-t}))$. Based on equation (2), it can be shown by some calculation,

$$s(\mathbf{x}) = 2(D-1)\mathrm{trace}(\mathbb{\Sigma}) - 2h(\mathbb{\Sigma}), \quad (3)$$

and

$$s(\mathbf{x}_{-t}) = 2(D-2)\mathrm{trace}(\mathbb{\Sigma}_{-t}) - 2h(\mathbb{\Sigma}_{-t}), \quad (4)$$

where $h(\mathbf{z})$ is the sum of the off-diagonal elements of a square matrix $\mathbf{z}$ and $\mathbb{\Sigma}_{-t}$ is the matrix $\mathbb{\Sigma}$ removing its tth row and tth column. Then combining (3) and (4) we can have

$$\frac{s(\mathbf{x})}{2(D-1)} - \frac{s(\mathbf{x}_{-t})}{2(D-2)} = \sigma_{tt} - \frac{1}{D-1}h(\mathbb{\Sigma}) + \frac{1}{D-2}h(\mathbb{\Sigma}_{-t}). \quad (5)$$

Similarly, choosing $r \neq t$, we can have

$$\frac{s(\mathbf{x}_{-r})}{2(D-2)} - \frac{s(\mathbf{x}_{-\{r,t\}})}{2(D-3)} = \sigma_{tt} - \frac{1}{D-2}h(\mathbb{\Sigma}_{-r}) + \frac{1}{D-3}h(\mathbb{\Sigma}_{-\{r,t\}}), \quad (6)$$

where $\mathbf{x}_{-\{r,t\}}$ is the variable $\mathbf{x}$ excluding the rth and tth elements, and $\mathbb{\Sigma}_{-\{r,t\}}$ is the corresponding basis covariance matrix. Subtracting (6) from (5), we obtain

$$\begin{aligned}\frac{s(\mathbf{x})}{2(D-1)} - \frac{s(\mathbf{x}_{-t})}{2(D-2)} - \frac{s(\mathbf{x}_{-r})}{2(D-2)} + \frac{s(\mathbf{x}_{-\{r,t\}})}{2(D-3)} \\ = -\frac{1}{D-1}h(\mathbb{\Sigma}) + \frac{1}{D-2}h(\mathbb{\Sigma}_{-t}) \\ + \frac{1}{D-2}h(\mathbb{\Sigma}_{-r}) - \frac{1}{D-3}h(\mathbb{\Sigma}_{-\{r,t\}}).\end{aligned} \quad (7)$$

Note that the right hand side of equation (7) is a linear combination of off-diagonal elements of $\mathbb{\Sigma}$ whereas the left hand side can be estimated using log ratios of the observed data. Without loss of generality, let us assume that $\boldsymbol{v} = [\sigma_{1D}, \sigma_{2D}, \ldots, \sigma_{D-1,D}, \sigma_{1,D-1}, \ldots, \sigma_{13}, \sigma_{23}, \sigma_{12}]^T$ be a $\frac{D(D-1)}{2} \times 1$ vector whose elements are the upper diagonal part of $\mathbb{\Sigma}$ and arranged in this particular order. Then, we can rewrite (7) as

$$w_{(r,t)} = \boldsymbol{a}_{(r,t)}^T \boldsymbol{v}, \quad (8)$$

where $w_{(r,t)}$ is the left hand side of (7) and $\boldsymbol{a}_{(r,t)}^T$ is a $\frac{D(D-1)}{2} \times 1$ vector of coefficients of $\boldsymbol{v}$ in the right hand side of (7). Note that while $w_{(r,t)}$ depends on data $\mathbf{x}$, the coefficients $\boldsymbol{a}_{(r,t)}^T$ depend only on the total number of taxa D, and the choices of r and t.

Since there are $\binom{D}{2} = \frac{D(D-1)}{2}$ choices of r and t, we can construct a linear system with exactly $\frac{D(D-1)}{2}$ unique equations from (8). For all $\frac{D(D-1)}{2}$ possible combinations of pairs of D compositions/taxa, let $\boldsymbol{w} = [w_{(D,1)}, w_{(D,2)}, \ldots, w_{(D,D-1)}, w_{(D-1,1)}, \ldots w_{(2,1)}]^T$, then we can have

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{a}_{(D,1)}^T \\ \boldsymbol{a}_{(D,2)}^T \\ \vdots \\ \boldsymbol{a}_{(D,D-1)}^T \\ \boldsymbol{a}_{(D-1,1)}^T \\ \vdots \\ \boldsymbol{a}_{(2,1)}^T \end{bmatrix} \boldsymbol{v} = A\boldsymbol{v}. \quad (9)$$

Since equation (7) is derived from the rank-deficient system based on (1), the linear system (9) has a rank of $\frac{D(D-1)}{2} - D$ with $\frac{D(D-1)}{2}$ unknown variables. We obtain the solution to (9) by introducing l_1-norm penalization

$$\min_{\boldsymbol{v}} \frac{1}{2}\|A\boldsymbol{v} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{v}\|_1, \quad (10)$$

where λ is a tuning parameter controlling the amount of non-zero solutions in $\boldsymbol{v}$. The penalized least-square method (10) is well-known as the LASSO (Tibshirani, 1996).

Note that a rank-deficient system usually has infinite solutions, but we show that the system (9) has a unique solution under certain sparse condition (Supplementary file).

2.1.2 Accessing significance

Since the above penalized solution (10) does not give proper P-values for selected variables, to access the significance of the selection we use a stability resampling method (Shah and Samworth, 2013). This method controls the expected number of selected variables with low selection probability, or family wise errors (FWE), while achieving higher power than the original stability selection method (Meinshausen and Bühlmann, 2010). The stability selection method derives a stability score τ for each variable based on the frequency at which the variable is being selected over a number of times. To be specific, we randomly split samples into two datasets B times, apply LASSO independently on the $2B$ datasets, and then obtain each solution and calculate the ratio of average number of selected over the total variables ρ. To control the rate of FWE at α (i.e. given data of D taxa, the expected number of low selection probability variables being selected is $\frac{D(D-1)}{2}\alpha$), we choose the minimum τ_α for selecting a variable based on

$$\begin{aligned}\tau_\alpha = \min_\tau \Bigg\{ \tau \in \Big\{0, \frac{1}{2B}, \ldots, 1\Big\} \\ : \min\Big\{ g\Big(\rho^2, 2\tau - 1, B, -\frac{1}{2}\Big), g\Big(\rho, \tau, 2B, -\frac{1}{4}\Big)\Big\} \leq \alpha \Bigg\},\end{aligned} \quad (11)$$

where g is a function without explicit form but can be evaluated numerically.

2.1.3 Algorithm

REBACCA can be summarized into the following steps:

(1) Input: count or proportion data $X \in \mathbb{R}^{D \times n}$ for D taxa
   1. Construct matrix A and compute $\boldsymbol{w}$ as in (9).
   2. Compute LASSO path Λ for (10).

(2) For $k = 1$ to B do
   1. Randomly split samples into two parts $\{X_1^{(2k-1)}, X_2^{(2k)}\} \subset \mathbb{R}^{D \times \frac{n}{2}}$.
   2. Compute $\boldsymbol{w}_1^{(2k-1)}$ and $\boldsymbol{w}_2^{(2k)}$ based on the random samples.
   3. Solve for $\boldsymbol{v}_1^{(2k-1,\lambda)}$ from $\left(A, \boldsymbol{w}_1^{(2k-1)}\right)$, and for $\boldsymbol{v}_2^{(2k,\lambda)}$ from $\left(A, \boldsymbol{w}_2^{(2k)}\right)$ using LASSO for each tuning parameter $\lambda \in \Lambda$.

(3) Assess significance
   1. Calculate selection frequency $F_{ij}^\lambda = \sum_{k=1}^{2B} |sign\left(\sigma_{ij}^{(k,\lambda)}\right)|$. Obtain the maximum frequency $F_{ij} = \max F_{ij}^\lambda$ over the LASSO path and the stability score $\tau_{ij} = \frac{F_{ij}}{2B}$.
   2. Choose a cutoff τ_α to control for FWER at α based on (11).

(4) Obtain estimation for $\mathbb{\Sigma}$
   1. Obtain $\{\sigma_{ij} : \tau_{ij} > \tau_\alpha\}$ and solve (9) by the least-square fit with the remaining variables constrained to be zero.
   2. Calculate diagonal elements of $\mathbb{\Sigma}$ according to equation (1).

2.2 Methods for generating compositional data

To simulate a metagenomic compositional data set, we consider the count data are drawn with two steps. We first generate basis abundance and proportion for each taxon, and then generate count data given a sequencing size. To be specific, the first step is called 'basis sampling'. We assume that basis proportions vary from sample to sample, and they are generated from one of three different underlying distributions, namely, log ratio normal (LRN), Poisson log normal and Dirichlet log normal distributions (Supplementary file). The second step is called 'sequencing sampling'. Count data are drawn from a multinomial distribution using the proportions obtained from the first step and a given total number of reads (i.e. library size) as parameters. The second step of sampling reflects a random process that all sequences are equally likely to be selected in a microbial community sample.

For LRN, given mean basis abundance $\boldsymbol{b} = (b_1, b_2, \ldots, b_D)^T$ and basis covariance $\mathbb{\Sigma}_{D\times D}$ for D taxa, we sample $\varphi_i = log\left(\frac{x_i}{x_D}\right) \sim MVN(\mu, \Omega)$ for $i = 1, \ldots, D-1$, where $\mu = \log\left(\frac{(b_1, b_2, \ldots, b_{D-1})^T}{b_D}\right)$, $\Omega = L\mathbb{\Sigma}L^T$, and $L = \begin{pmatrix} 1 & \cdots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1 \end{pmatrix}_{(D-1)\times D}$. Then, we can obtain basis proportions $\omega_i = \frac{\exp(\varphi_i)}{1+\sum_{i=1}^{D-1}\exp(\varphi_i)}$ for $i = 1, \ldots, D-1$, and $\omega_D = \frac{1}{1+\sum_{i=1}^{D-1}\exp(\varphi_i)}$. In the second step, the count data are drawn from $x_1, x_2, \ldots, x_D \sim \begin{pmatrix} x_+ \\ \boldsymbol{x} \end{pmatrix} \prod_{i=1}^{D}\omega_i^{x_i}$ for each sample given a sequencing size of $x_+ = \sum_{i=1}^{D} x_i$.

2.3 Metagenomic dataset

We use a real metagenomic dataset from a previous study on mouse skin microbiota (Srinivas *et al.*, 2013). The dataset contains 131 core OTUs. We analyze their abundance data from 261 mouse skin samples, including 78 non-immunized and 183 immunized individuals (of which 64 developed epidermolysis bullosa acquisita or EBA, a skin blister disease). The most abundant phylum is Firmicutes (44 OTUs), accounting for 49% of abundance on average, followed by Proteobacteria (35 OTUs), Bacteroidetes (26), Actinobacteria (17), Cyanobacteria (4) and others (5).

3 Simulation studies

We generate synthetic samples under different conditions. We compare results obtained from REBACCA with other three methods including the conventional correlation method [bootstrapping method (BP)], ReBoot (Faust *et al.*, 2012) and SparCC (Friedman and Alm, 2012), in terms of their sensitivity and specificity. Results show that while REBACCA is highly efficient as compared with SparCC, it outperforms other methods with its high sensitivity as well as its control for false positives.

3.1 Simulation strategy

To evaluate REBACCA comprehensively, we consider that count data are generated from basis abundance with three cases of covariance structures (Supplementary Fig. S1). The three cases represent three different sparse network structures, such that Case 1 has a hierarchical structure, Case 2 has a group of four highly connected taxa and Case 3 is larger (20 taxa) than the first two cases (10 taxa for each case) and has three independent sub-network groups. Besides the three cases, we also simulate data for 100 taxa (see Section 3.3). However, we mainly focus on the small-scale synthetic data for two reasons. First, the issue of artifactual correlation tends to be more severe with smaller number of components (Friedman and Alm, 2012). Second, inaccuracy in estimating correlations can be due to low strength of signals or small magnitudes of correlations, which can obscure the results when dealing with artifactual correlation. In each of Case 1, 2 and 3, the magnitudes of correlations for the correlated pairs ranged from 0.3 to 0.6.

Besides the structures of basis covariance, we also consider two types of distributions of counts over the taxa, such as whether all taxa are 'equal' or 'unequal' in terms of their mean basis abundances. In a real dataset, counts are usually not distributed evenly for different taxa, and as a consequence the effect of spurious correlations cannot be ignored even on datasets with large number of taxa (Friedman and Alm, 2012).

In each of the six small-scale situations (3 covariance cases × 2 mean types), we simulate different sizes of microbiome samples ($n = 50,\ 100,\ 200$), with the total number of counts equal to 3000 per sample. Finally, to compare the results from different methods, we simulate each situation 100 times.

3.2 Results on small-scale studies

We compare REBACCA with other three methods including the Spearman's correlation with BP, ReBoot and SparCC. All these methods assess significance of pair-wise correlations based on a resampling size of 100. For REBACCA, we use one minus stability score as the pseudo *P*-value.

We obtain the receiver operating characteristic (ROC) curve based on the *P*-values of all pair-wise correlations for each simulation result, and summarize it with the area under the curve (AUC). In general, higher AUC indicates that the corresponding method has overall higher ratio of power over Type I errors. For the synthetic datasets generated using LRN method with a sample size of 100, the medians of AUC obtained by REBACCA are higher than all other methods in Case 1 and 3, whereas they are at least 0.875 in Case 2 (Fig. 1).

SparCC and ReBoot have higher AUC than BP in Case 1 and 3, but lower in Case 2. It is possible that the accuracies of these methods depend on whether the estimated correlations are positive or negative, because there are more negative correlations in Case 2 than other two cases (Supplementary Fig. S1). We obtain the mean ROC curve separately for positive and negative correlations in Case 1, and we find that the accuracies of ReBoot and SparCC are comparable to REBACCA when identifying positive correlations but substantially lower when only considering the negative ones (Supplementary Fig. S2). BP has the lowest accuracy of all when identifying positive correlations.

When calculating correlations between components with unequal proportions, conventional methods can lead to incorrect inference on their relationships (Lovell *et al.*, 2010). We show that

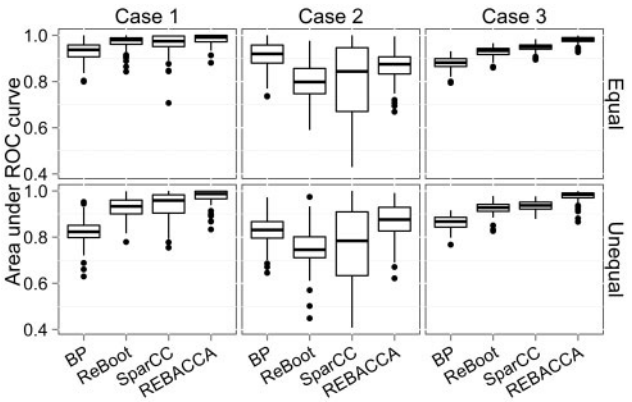

Fig. 1. Comparison of methods based on AUC. Each boxplot represents the AUC values calculated on 100 simulated datasets. Data are generated using LRN method with 100 samples based on three cases of structures of basis covariance including 'Case 1' with a hierarchical structure, 'Case 2' consisting four inter-correlated taxa and mostly negative correlations and 'Case 3' with three clustered groups. Two types of mean basis abundance are used in simulation such that average basis abundance are 'equal' or 'unequal' for different OTUs. Four methods are compared including the conventional correlation measure with resampling (BP), ReBoot,SparCC and REBACCA

correlations measured by BP and ReBoot are not consistent between two types of mean basis abundance where counts are more evenly distributed over taxa in one than the other (Fig. 1). REBACCA performs consistently well on different types of mean distributions.

Given larger sample size, REBACCA always yields higher AUC (Supplementary Fig. S3), while in Case 3 other three methods do not gain better accuracies when samples are more than 100.

We also compare these methods based on their error controls. Given a pre-specified significance level of 0.05, only REBACCA controls false positives consistently at a rate <0.05 in all situations (Fig. 2). BP and ReBoot result in larger number of errors than targeted in all situations. SparCC controls errors well in Case 1 but fails when many negative correlations are present as in Case 2 or a large number of pairs (45) are correlated as in Case 3. Remarkably, REBACCA can identify correlated pairs while controlling the minimum false positive rates (FPR) lower than 0.005 in all situations (Fig. 3). In Case 1 and 3, it identifies correctly more than 70% correlated pairs with FPR <0.01. Other three methods can identify correlations only when FPR is allowed higher than 0.01 in most situations.

The conclusions from these simulation studies are summarized as follows. First, REBACCA has higher accuracy in identifying correlated taxa in five of the six situations, while offering the best control of FPR (Fig. 3). Second, its estimated correlation structure converges to the true one provided larger number of samples (Supplementary Fig. S3). Third, it has the most consistent performance regardless of equal or unequal mean situations, and whether the estimated correlations are positive or negative. Fourth, note that although count data can be simulated through different schemes rather than LRN, we obtain similar results on datasets generated by other two methods (Supplementary Fig. S4–S5).

3.3 Results on large-scale studies

We examine REBACCA and other methods on data generated with 100 OTUs and two covariance structures, which are based on a scale-free network model and a clustered network with three subnetworks (similar to Case 3). While the average AUC for REBACCA is only slightly higher than other methods for the scale-free network model, it is substantially higher in the clustered case (Supplementary Fig. S6). Remarkably, REBACCA gains higher AUC for the clustered network when sample size increases from 100 to 200 (with almost exact recovery at a sample size of 200), whereas other methods fail to achieve better accuracy. It is possible that the solutions by BP, ReBoot and SparCC depend on either mitigating the compositional bias or compensating the approximation rather than exact estimation of the correlation structure. The results for the scale-free and clustered models indicate that REBACCA is also promising when analyzing large datasets.

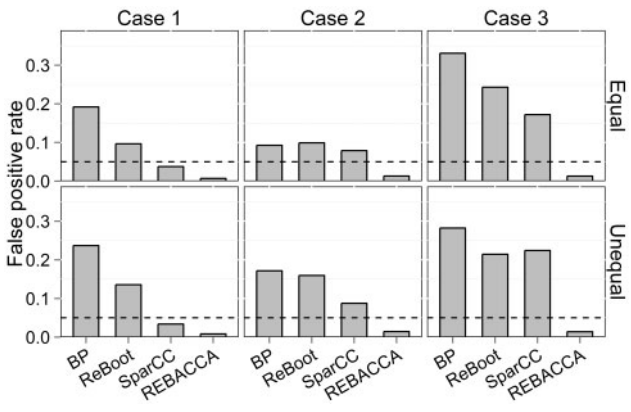

Fig. 2. Comparison of methods based on their controls for Type I error. Dash lines represent a targeted FPR at 0.05. That is, we consider a pair is correlated if the corresponding *P*-value is <0.05.Bars represent mean FPRs for 100 simulated datasets for each situation. Data are generated using LRN method with 100 samples

4 Analysis of mouse skin data

We apply REBACCA to identify pair-wise correlations on a mouse skin data including three groups of individuals: groups of non-immunized (Control), immunized-healthy (Healthy) and immunized-diseased (EBA) individuals. To assess the significance and control the error rate, we randomly split samples into two datasets, identify their non-zero basis covariance independently, and merge the results (Section 2). We do this 500 times, and thus the estimated covariance between a pair of OTUs is assigned with a stability score. Correlated pairs are identified based on stability scores with cutoffs of 0.954, 0.951 and 0.954 for the Control, Healthy and EBA groups, respectively. These cutoffs correspond to a FWER of 0.03, that is, the expected number of falsely identified correlated pairs is <255 out of 8515 possible pairs. Different cutoffs for stability scores are also chosen to verify our results.

4.1 Within-phylum correlations are more common

Based on the Control group, we identify 606 pair-wise correlations among the 131 core OTUs controlling the FWER at 0.03. The largest number of taxa correlated with a taxon is 40 (Supplementary Fig. S7), whereas on average a taxon correlates with 9.3 taxa. There are 416 positively correlated pairs and 190 negative ones.

We first investigate whether our results support the finding from a previous study that phylogenetically related species are more likely to be co-occurred in their abundances, whereas the diverse species are more likely to be co-excluded (Chaffron *et al.*, 2010). We group the OTUs based on four major phyla (Firmicutes, Proteobacteria,

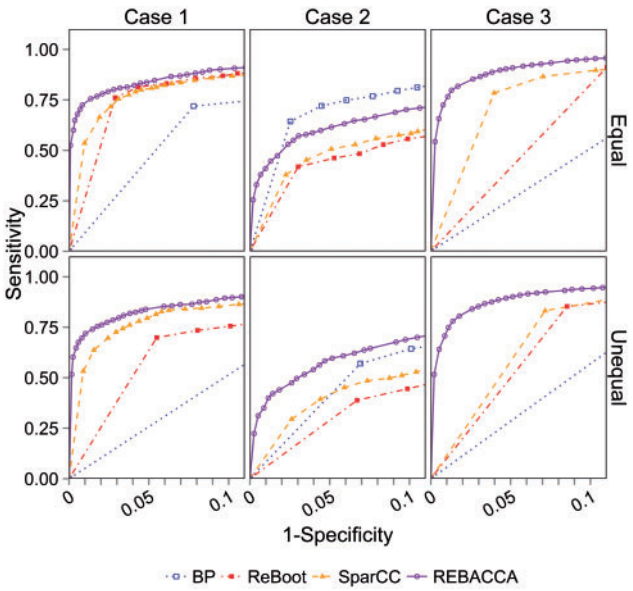

Fig. 3. Comparison of methods based on their powers (sensitivity) and FPRs (1-specificity). A sequence of cutoffs with intervals of 0.01 is used to calculate the sensitivity and specificity. Points represent the average of the results for 100 simulated datasets in each situation. Data are generated using LRN method with 100 samples

Bacteroidetes and Actinobacteria), and compare the amount of associated OTUs within each phylum and between the phyla using the fraction of identified associations over the total number of possible pairs. Among the 122 OTUs from the major four phylum clades, there are 253 pairs of associations or 13% of all possible pairs within the same phylum (Fig. 4a and Table 1), whereas there are 354 associations or 3% of all possible pairs between different phyla. Within each clade, there are much more positive associations than the negative ones, whereas between different clades the difference between number of positive and negative ones is not substantial (Table 1). Together the large proportion of OTU associations and the amount of positive ones within each phylum clade suggest that co-occurrence relationships are more common in closely related species than those are distantly related.

We are also interested in whether the co-occurrence patterns indicate ecological relationships between bacteria species. A biofilm is usually formed by a group of interacting bacteria and can protect its members from hostile environments (Hall-Stoodley *et al*., 2004), and thus it may exhibit co-existing relationships between the members. Biofilms formed by *Staphylococcus* species on skins have been known for multidrug resistance (Leroy *et al*., 2009; Otto, 2008), and biofilms formed by *Corynebacterium* species can potentially cause infections (Kwaszewska *et al*., 2006). Our result shows that among top 61 (∼10%) strongest positively correlated pairs of OTUs, 38 are pairs between *Staphylococcus* species and 6 are between *Corynebacterium* species. Another interesting relationship is that *Staphylococcus aureus* and *Enterococcus faecalis* are positively correlated (∼0.34), reflecting the fact that the former species can secrete a sex pheromone specific to the latter one and thus promoting the reproduction of *E. faecalis* (Muscholl-Silberhorn *et al*., 1997).

To compare our result with the correlated OTUs identified by SparCC (Friedman and Alm, 2012), we choose the significance level at 0.003 for SparCC, so that we obtain a comparable number of correlated pairs (608 as compared with 606 by our method). Under SparCC, there are 369 positively correlated pairs and 239 negative ones. Notably, there are only 91 (∼25%) positive correlated pairs identified by SparCC that are considered uncorrelated by REBACCA, whereas 133 (∼56%) negative ones do not agree with the result from REBACCA (Supplementary Table S2). This may be due to the lower accuracy of estimation by SparCC on negative correlations than the positive ones (Supplementary Fig. S2). Nevertheless, SparCC agrees with REBACCA on that co-occurrence relationships are more common within the same phylum than between phyla (Supplementary Table S3).

4.2 Correlation patterns are more similar between two types of immunized samples

We investigate whether the correlation patterns among the 131 OTUs are different depending on the three types of samples, namely, Control, Healthy and EBA. We compare correlated pairs that are identified independently from the three datasets, and we consider that a correlated pair of OTUs is consistent between two datasets if the pair has the same signs of correlations in both datasets. We find that correlations from the non-immunized individuals are much less consistent with other two immunized groups than between the two groups (Fig. 4). Although there are 532 consistent pairs between the two immunized groups, there are only 236 consistent pairs between the Control and Healthy and 212 between the Control and EBA groups (Fig. 4d).

The similarity of the correlation patterns between two immunized groups indicates a common change in skin microbiota due to immunization. A network consisting of positive correlations among 12 Bacteroidia of phylum Bacteroidetes and 6 Clostridia of Firmicutes is found in both immunized groups whereas absent in the Control group (Fig. 4). Bacteroidia and Clostridia are anaerobic bacteria known for their antibiotic resistance (Bryan *et al*., 1979; Wexler, 2007). Positive correlations among these bacteria may promote their survivals and cause polymicrobial infections to their host (Brook *et al*., 1984).

Although the number of identified correlations depends on the choice of the FWER control, it is worth noting that regardless using more or less stringent choices (FWER = 0.01, 0.03 and 0.05), the patterns of dominant positive within-phylum correlations and the presence of inter-phyla correlation patterns exclusively in the immunized individuals are consistent on all these results (Supplementary Figs S8 and S9).

The pattern of positive correlations among the Bacteroidetes and Clostridia is also observed in the results obtained by SparCC (Supplementary Fig. S10), although the correlations in the two immunized groups are weaker than those estimated by REBACCA.

5 Discussion

We propose a method to study pair-wise microbial relationships for metagenomic data. Applying conventional measures directly on proportion data can lead to incorrect inference on the pair-wise relationship due to spurious correlation. By assuming the underlying basis abundance, REBACCA infers pair-wise relationships of basis abundance and thus is free of compositional bias. In general, it provides higher power on identifying correlated pairs than other existing methods while controlling false positives.

Another major advantage of REBACCA as compared with SparCC is the computational efficiency. SparCC solves the system (2) relying on recursively refining its solution by identifying strong correlations, and its computing time and accuracy depends on the number of strongly correlated pairs. The computing time for REBACCA does not depend on the complexity of the data and is nearly independent of the size of samples. For the mouse skin

Table 1. Fraction of correlated pairs of OTUs over the total

Phylum	Within-phylum		With other phyla	
	Fraction of correlated over possible pairs (positive, negative)	Number of correlated pairs	Fraction of correlated over possible pairs (positive, negative)	Number of correlated pairs
Actinobacteria	0.103 (0.096, 0.007)	14	0.078 (0.05, 0.029)	140
Bacteroidetes	0.046 (0.040, 0.006)	15	0.024 (0.014, 0.01)	60
Firmicutes	0.169 (0.147, 0.022)	160	0.06 (0.03,0.031)	207
Proteobacteria	0.108 (0.096, 0.012)	64	0.051 (0.026, 0.025)	155

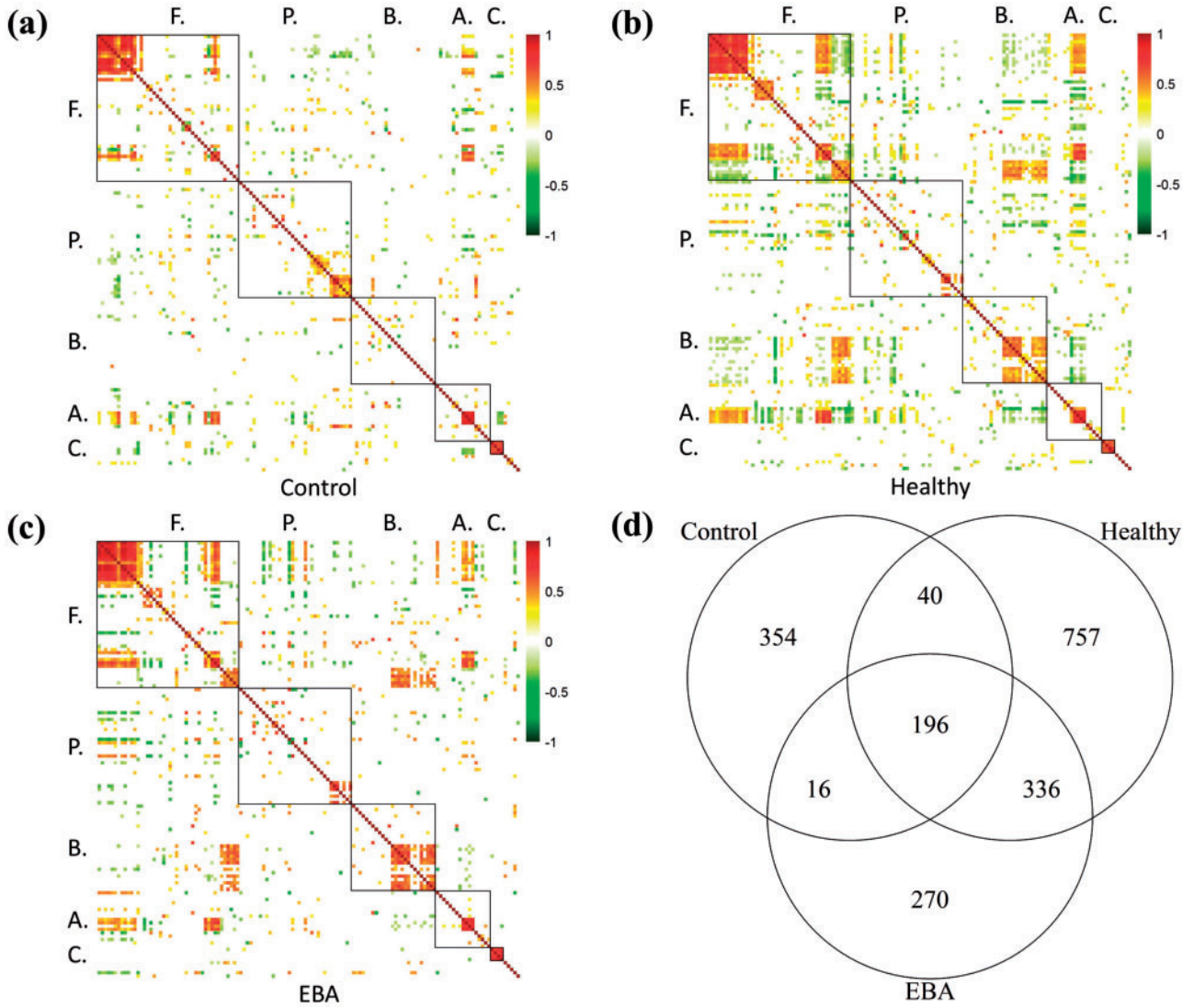

Fig. 4. Basis correlations between core OTUs. Results of correlations calculated from three types of mouse skin microbiota samples are shown, including (**a**) 606 correlated pairs from non-immunized (Control), (**b**) 1329 pairs from immunized but healthy (Healthy), (**c**) 818 pairs from immunized and developed EBA disease (EBA) samples. Within-phylum correlations are shown in the square areas for Firmicutes (F.), Proteobacteria (P.), Bacteroidetes (B.), Ctinobacteria (A.) and Cyanobacteria (C.). Correlated pairs are identified with FWER controlled at 0.03. (**d**) Venn diagram of consistent correlated OTUs from Control, Healthy and EBA samples. There are significantly more correlated pairs consistent between the immunized groups than other comparisons

Control dataset, running SparCC with 1000 bootstraps on a cluster of 20 cores of Intel Westmere X5650 processors (2.66 GHz, 4 GB memory per core) requires more than 10.5 h, whereas running REBACCA with $B = 500$ (equivalent to 1000 bootstraps) on the same cluster takes only about 30 min. A fast implementation of SparCC reduces the computing time (Kurtz *et al.*, 2015). Nevertheless, both the original and improved version of SparCC require setting a maximum number of recursions (the default is 10), which can reduce their computing time but also lower their accuracy when there are quite a few strong correlations between OTUs such as the clustered cases (Fig. 1 and Supplementary Fig. S6).

We show that the rank-deficient system (9) is identifiable or can be solved exactly, provided a sparse condition that each taxon correlates with less than a quarter of the total number of taxa (Supplementary file). In the case where a component is positively correlated with all other components with the same magnitude, none of these correlations can be identified no matter how large the sample size is. However, based on our experiments, if some correlations are positive and some are negative, REBACCA can find the exact solution despite that the sparse condition is violated. Above all, the sparse condition is derived from an ideal, noise-free case, and thus in reality, even if the sparse condition is satisfied, there still can be an issue of identifiability for the system in (9). This is because that given small sample size, correlation can be substantially influenced by uncertainties. The noise and identifiability issues may be the main reason that we obtain lower AUC for REBACCA in Case 2 than other two cases (Fig. 2). Nevertheless, it should be noted that the sparse condition applies to the model in (9), which is equivalent to (2) and thus the condition should also apply to SparCC. In the high-dimensional setting of a real metagenomic dataset, the sparse condition should usually hold, because bacterial interaction network is highly sparse (Freilich *et al.*, 2010).

Although it is possible to solve the system (2) with some regularization such as constraining the diagonal elements of Σ to be positive, the error distribution in the linear system is unlikely to be normal and thus further transformation may be required. The

construction of the linear system in (9) is not unique; however, the system constructed by equation (7) is optimal because the right hand side of (7) contains information of all variables, and the homogeneity assumption of error variance is satisfied if the distributions for upper diagonal variables in Σ are exchangeable.

Calculation of *P*-values in high-dimensional variable selection methods is an open problem. Stability selection offers a way to control for FWER using a sample-splitting resampling method (Meinshausen and Bühlmann, 2010), however, based on our experiments we find that this method is too conservative for the mouse skin data. Instead, we use the modified stability selection method proposed by Shah and Samworth (2013). This method has much better power while also controlling for FWER.

Change of bacteria abundance in mouse skin microbiota is linked to the cause of skin disease (Srinivas *et al*., 2013). We show that the microbial correlation patterns in immunized samples are different from the non-immunized ones. The difference highlights the response of a group of Bacteroidetes and Clostridia to immunization (Fig. 4). These species have been reported associated with anaerobic infections. Further analysis of the difference of bacterial correlations between the healthy and diseased groups may reveal the link between microbial factors and disease susceptibility.

Funding

This work was supported by National Science Foundation [DMS-1043080 to H. J. and L.A.] and [DMS-1222592 to H.J. and L.A.], and partially supported by National Institutes of Health [P30 ES006694 to L.A.] and USDA National Institute of Food and Agriculture, Hatch project [ARZT-1360830-H22-138 to L.A].

Conflict of Interest: none declared.

References

Aitchison,J. (1981) A new approach to null correlations of proportions. *J. Int. Assoc. Math. Geol.*, **13**, 175–189.

Brook,I. *et al*. (1984) Synergistic effect of bacteroides, Clostridium, Fusobacterium, anaerobic cocci, and aerobic bacteria on mortality and induction of subcutaneous abscesses in mice. *J. Infect. Dis.*, **149**, 924–928.

Bryan,L.E. *et al*. (1979) Mechanism of aminoglycoside antibiotic resistance in anaerobic bacteria: *Clostridium perfringens* and *Bacteroides fragilis*. *Antimicrob. Agents Chemother.*, **15**, 7–13.

Chaffron,S. *et al*. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.*, **20**, 947–959.

Chen,J. and Li,H.Z. (2013) Variable selection for sparse Dirichlet-Multinomial regression with an application to microbiome data analysis. *Ann Appl Stat.*, **7**, 418–442.

Egozcue,J.J. *et al*. (2003) Isometric logratio transformations for compositional data analysis. *Math Geol.*, **35**, 279–300.

Faust,K. *et al*. (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.*, **8**, e1002606.

Freilich,S. *et al*. (2010) The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.*, **38**, 3857–3868.

Friedman,J. and Alm,E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, **8**, e1002687.

Hall-Stoodley,L. *et al*. (2004) Bacterial biofilms: from the natural environment to infectious diseases. *Nat. Rev. Microbiol.*, **2**, 95–108.

Hojo,K. *et al*. (2009) Bacterial interactions in dental biofilm development. *J. Dent. Res.*, **88**, 982–990.

Katsuyama,C. *et al*. (2009) Complementary cooperation between two syntrophic bacteria in pesticide degradation. *J. Theor. Biol.*, **256**, 644–654.

Kwaszewska,A.K. *et al*. (2006) Hydrophobicity and biofilm formation of lipophilic skin corynebacteria. *Pol. J. Microbiol.*, **55**, 189–193.

Kurtz,Z.D. *et al*. (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, **11**, e1004226.

Leroy,S. *et al*. (2009) Genetic diversity and biofilm formation of Staphylococcus equorum isolated from naturally fermented sausages and their manufacturing environment. *Int. J. Food Microbiol.*, **134**, 46–51.

Lovell,D. *et al*. (2010) Caution! compositions! technical report and companion software (publicationtechnical). Technical Report EP10994, CSIRO.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **72**, 417–473.

Muscholl-Silberhorn,A. *et al*. (1997) Why does *Staphylococcus aureus* secrete an *Enterococcus faecalis*-specific pheromone?. *FEMS Microbiol. Lett.*, **157**, 261–266.

Orphan,V.J. *et al*. (2001) Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. *Science*, **293**, 484–487.

Otto,M. (2008) Staphylococcal biofilms. *Curr. Topics Microbiol. Immunol.*, **322**, 207–228.

Shah,R.D. and Samworth,R.J. (2013) Variable selection with error control: another look at Stability Selection. *J. R. Stat. Soc. Ser. B*, **75**, 55–80.

Srinivas,G. *et al*. (2013) Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.*, **4**, 2462.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B Methodol.*, **58**, 267–288.

Wexler,H.M. (2007) Bacteroides: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.*, **20**, 593–621.