BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 158–165
doi:10.1093/bioinformatics/btu635

Genome analysis

Advance Access publication September 24, 2014

A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes

Naruekamol Pookhao[1], Michael B. Sohn[2], Qike Li[2], Isaac Jenkins[2], Ruofei Du[1], Hongmei Jiang[3] and Lingling An[1,2,*]

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

Associate Editor: John Hancock

ABSTRACT

Motivation: With the advance of new sequencing technologies producing massive short reads data, metagenomics is rapidly growing, especially in the fields of environmental biology and medical science. The metagenomic data are not only high dimensional with large number of features and limited number of samples but also complex with a large number of zeros and skewed distribution. Efficient computational and statistical tools are needed to deal with these unique characteristics of metagenomic sequencing data. In metagenomic studies, one main objective is to assess whether and how multiple microbial communities differ under various environmental conditions.

Results: We propose a two-stage statistical procedure for selecting informative features and identifying differentially abundant features between two or more groups of microbial communities. In the functional analysis of metagenomes, the features may refer to the pathways, subsystems, functional roles and so on. In the first stage of the proposed procedure, the informative features are selected using elastic net as reducing the dimension of metagenomic data. In the second stage, the differentially abundant features are detected using generalized linear models with a negative binomial distribution. Compared with other available methods, the proposed approach demonstrates better performance for most of the comprehensive simulation studies. The new method is also applied to two real metagenomic datasets related to human health. Our findings are consistent with those in previous reports.

Availability: R code and two example datasets are available at http://cals.arizona.edu/~anling/software.htm

Contact: anling@email.arizona.edu

Supplementary information: Supplementary file is available at *Bioinformatics* online.

Received on May 1, 2014; revised on September 5, 2014; accepted on September 19, 2014

1 INTRODUCTION

Recently next-generation sequencing technologies are able to produce high volumes of data at an affordable cost (Gilbert *et al.*, 2011; Huson *et al.*, 2009). The power of next-generation sequencing makes it possible to explore microbial environments, opening a new era of genomics study, called metagenomics (Gilbert *et al.*, 2011). Metagenomics is the study of genomic contents of microbial communities sampled directly from environments (e.g. soil, water, human gut) without prior culturing to understand the true diversity of microbes, their functions, cooperation and evolution in different microbial communities (Hugenholtz, 2002; Huson *et al.*, 2009; Kunin *et al.*, 2008; Wooley and Ye 2010). Importantly, because only ~1% of all microbial organisms can be isolated and cultured in a laboratory, metagenomic analysis enables to reveal the genome contents of the majority of microorganisms that cannot be obtained in traditional genomic analysis based on pure culture (Hugenholtz, 2002; Wooley and Ye, 2010). Metagenomics is broadly applicable to many areas, including ecology and environmental sciences, chemical industry and biomedicine (Turnbaugh *et al.*, 2007; Wooley and Ye, 2010).

In metagenomic analysis, one important aim is to assess whether and how two or more microbial communities differ. To perform metagenomic comparison, researchers can conduct an experiment to compare genomic features based on either taxonomic compositions or functional components obtained from different microbial communities. In this study, we focus on comparison of functions in metagenomes under various conditions. The applications of this research include detection of biological threats and discovery of new bioenergy and new medicine, and so on. For example, comparing microbial communities from human gut corresponding to different phenotypes (e.g. diseased and healthy, or different treatments) can help us determine the activities of microbes related to the disease, resulting in understanding the reactions of microbes that respond to different biochemical products. This may lead to drug development or treatment selection that specifically affects either a particular activity or a group of activities that the disease-related microbes might perform.

Statistical procedures play a critical role in detecting differentially abundant features across different microbial conditions. The features here may refer to taxa, functional roles, pathways, or subsystems. Several statistical methods or tools have been developed to compare various microbial communities in terms of detecting differentially abundant features, e.g. SONs (Schloss and Handelsman, 2006), XIPE-TOTEC (Rodriguez-Brito *et al.*, 2006), Metastats (White *et al.*, 2009) and MEGAN (Huson *et al.*, 2009, 2011). However, all of these methods/tools are designed to compare exactly two microbial conditions; ShotgunFunctionalizeR uses a regression method on comparing multiple samples (Kristiansson *et al.*, 2009) but it assumes Poisson

*To whom correspondence should be addressed.

distribution on the count data. It is well known that Poisson model lacks flexibility for over-dispersed count data (Rapaport *et al.*, 2013). Another method, metagenomeSeq (Paulson *et al.*, 2013), has been recently developed to assess differential abundance in sparse high-throughput microbial marker-gene survey data. Even though it can compare multiple conditions, metagenomeSeq is designated for comparison of taxonomic compositions of different metagenomes, rather than functional compositions. In this research, we focus on statistical comparison of functions in metagenomes under various conditions.

Statistical methods developed for RNA-Seq analysis may be applicable to metagenomic analysis also, as both RNA-Seq experiments and metagenomic experiments use sequencing technologies and produce count data. A number of statistical tools have been developed for RNA-Seq data analysis, such as edgeR (Robinson *et al.*, 2010) and DESeq (Anders and Huber, 2010). However, there are differences between RNA-Seq data and metagenomic data. Different from RNA-Seq data, one of the common characteristics of metagenomic data is the presence of many features with zero counts. It is because metagenomic samples consist of a mixture of microbes, the species-specific functions may only appear in some microbial conditions, while in typical RNA-Seq experiments the genes are the same for different experimental conditions, and only expression levels change. Thus, metagenomic sequencing data may be more sparse than the RNA-seq data.

Our research was motivated by (i) the limitations of existing methods developed for metagenomic analysis, (ii) the increasing focus of metagenomic projects on wide applications in various areas [e.g. Human Microbiome Project (HMP, Turnbaugh *et al.*, 2007)] and (iii) the limitations of applying current methods developed for RNA-Seq analysis to metagenomic analysis. In this article, we propose a two-stage statistical algorithm for selecting informative features and detecting differentially abundant functional features (e.g. pathways, subsystems, functional roles) between different microbial conditions. In the first stage of our algorithm, the informative features are selected using elastic net (Friedman *et al.*, 2010) resulting in dimensional reduction of the metagenomic dataset. In the second stage of our approach, we detect differentially abundant features using generalized linear models (GLMs) with a negative binomial (NB) distribution (Venables and Ripley, 2002).

In sparse data, elastic net is a satisfactory variable selection method in the case that the number of predictors (p) is much bigger than the number of observations (N), that is, when $p >> N$. In addition, another advantage of elastic net is that it is well suitable to data containing a grouping effect, i.e. strongly correlated predictors tend to be in or out of the model together (Friedman *et al.*, 2010; Zou and Hastie, 2005). The NB distribution is widely used to model count data. The novelty of our two-step method is that we take the common characteristics of metagenomic data into account and combine the feature selection and feature comparison in metagenomic study to improve the power of feature detection.

Our method can be directly applied to comparison of more than two microbial conditions. Therefore, our method can be applicable to more general situations, e.g. in clinical trials where the goal is to compare multiple treatment conditions or in natural environmental studies where multiple conditions are compared and investigated.

2 METHODS

Our approach requires (i) a metagenomic dataset corresponding to two or more conditions/phenotypes (e.g. diseased and healthy human guts, or different locations of sea water); each condition/phenotype consists of multiple individuals (or samples), and (ii) each sample/individual consists of count data representing the relative abundance of features, or number of shotgun reads mapped to a specific biological pathway or subsystem. Our goals are to determine a set of informative features associated with a particular phenotype and to identify statistically significant features whose abundance is different among different conditions/phenotypes.

2.1 Data normalization

Due to the high-throughput sequencing technologies, an arbitrary number of reads with large variation across samples is generated under the sampling process. That is, a common source of bias in a metagenomic count data is owing to different sequencing depths or various magnitude of the read counts across multiple individuals (or samples). To proceed with any statistical analysis, a preprocessing of the metagenomic count data is necessary to account for this source of bias, i.e. normalizing the samples to make them comparable. For the data normalization, we used the trimmed mean of M-values (Robinson and Oshlack, 2010), which is implemented in the edgeR Bioconductor package.

2.2 Two-stage statistical procedure

In the proposed two-stage statistical algorithm, informative features are simultaneously selected in the first stage, and then the selected features obtained from the first stage are used as the input for the second stage. Differentially abundant features between metagenomic conditions/phenotypes are detected in the second stage.

First stage—feature selection using elastic net

The first stage aims to detect informative features associated with a particular phenotype. This results in the dimensional reduction of the metagenomic data. As outlined in the introduction, the metagenomic data consist of relative abundances where low abundant microorganisms may be missed owing to the sampling process. A statistical method is needed to deal with a sparse data with the presence of a large percentage of zero counts. Elastic net, an algorithm for estimation of GLMs with elastic-net penalties, enables to deal efficiently with sparse features (Friedman *et al.*, 2010).

For the first stage, assume there are p features and N samples. Let $a_s^T = [a_{1s}, a_{2s}, \ldots, a_{ps}]$ represent the vector of count values for p features in the sample s ($s = 1, \ldots N$), and the phenotype of sample s is denoted by g_s, which takes values from $\{1, 2, \ldots, K\}$ and K is the total number of phenotypes or categories. For example, when there are only two phenotypes (e.g. diseased and healthy), $K = 2$ and $g_s \in \{1, 2\}$.

Algorithm for elastic net

In a linear model, let G represent the response variable (e.g. phenotype status) and A represent the predictor variables, then the regression function is typically determined by $E(G|A = a) = \beta_0 + a^T\beta$, where a is a realization of the predictors. For N observation sets (a_s^T, g_s) $s = 1, \ldots, N$, the elastic net solves the following problem

$$\min_{(\beta_0,\beta)\in\Re^{p+1}} R_\lambda(\beta_0, \beta) = \min_{(\beta_0,\beta)\in\Re^{p+1}} \left[\frac{1}{2N}\sum_{s=1}^{N}(g_s - \beta_0 - a_s^T\beta)^2 + \lambda P_\alpha(\beta)\right] \quad (1)$$

where

$$P_\alpha(\beta) = \sum_{f=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_f^2 + \alpha|\beta_f| \right] \quad (2)$$

α is the elastic-net penalty (Zou and Hastie, 2005) and is a compromise between the ridge regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). The elastic net model with $\alpha = 1-\varepsilon$ for some small ε ($\varepsilon > 0$) performs much like the lasso, but ignores behavior caused by extreme correlations. This model will tend to pick one feature and ignore the rest if the features are correlated. On the other hand, the elastic net model with $\alpha = 1-\varepsilon$ for some large ε ($\varepsilon > 0$) performs much like the ridge regression, which is known as a regression model to shrink the coefficients of correlated predictor variables toward each other, resulting them to borrow strength from each other. The coordinate descent step used to solve (1) is detailed in Friedman *et al.* (2010).

Regularized multinomial regression

When the response variable is binary ($K = 2$), the linear logistic regression model is often used. When the categorical response variable G takes multiple values ($K > 2$), the linear logistic regression model can be generalized to a multi-logit model. The class-conditional probability is represented through a linear function of the predictors:

$$log\frac{Pr(G=\ell|a)}{Pr(G=K|a)} = \beta_{0\ell} + a^T\beta_\ell, \ell = 1, ..., K-1 \quad (3)$$

Here β_l is a p-vector of coefficients, and the parameters (βs) are computed by solving the penalized multinomial log-likelihood problem:

$$\max_{\{\beta_{0\ell},\beta_\ell\}_1^K \in \Re^{K(p+1)}} \left[\frac{1}{N}\sum_{s=1}^{N} log(Pr(g_s=\ell|a_s)) - \lambda\sum_{\ell=1}^{K} P_\alpha(\beta_\ell) \right] \quad (4)$$

where λ is a tuning parameter and will be determined as below.

Selecting the tuning α and λ parameters for regularization path

As shown in (1), two types of constraints (lasso and ridge constraints) on the parameters are used in the elastic net. The parameter α controls the relative weight of these constraints. The lasso constraints allow for the selection/removal of variables in the model, while the ridge constraints can deal with correlated predictor variables. In our approach, as the second step can deal with feature detection, in the elastic net step we put more weight on the ridge constraints to deal with correlated features. We use grid search for α in [0, 0.1], and for each parameter α, the corresponding λ was determined by cross-validation (CV) (Hastie *et al.*, 2009). The values for the parameters α and λ that yield the lowest CV error were selected.

Second stage—differentially abundant feature detection

The second stage of our algorithm is to detect features, which are statistically differentially abundant in two or more conditions. From examining real metagenomic count data, we discovered that the variance exceeds the corresponding mean of the feature abundance (detailed in Supplementary S1–S4). NB distribution, a commonly used model for count data with overdispersion, is used to take the overdispersion into account (Cameron and Trivedi, 1998; Venables and Ripley, 2002).

NB model

Assume r of p features are selected from the first stage. Let Y be the vector of the numbers of reads for feature i in all samples where $i = 1, 2, \ldots, r$. Each element (y_s) in Y can be modeled by NB distribution:

$$f_Y(y_s;\mu_s,\theta) = \frac{\Gamma(y_s+\theta)}{\Gamma(\theta)\cdot y_s!}\cdot\frac{\mu_s^{ys}\cdot\theta^\theta}{(\mu_s+\theta)^{y_s+\theta}} \quad (5)$$

with mean $E(y_s) = \mu_s$ and variance $var(y_s) = \mu_s(1 + \mu_s/\theta)$. The variance is quadratic in the mean. The NB distribution can also be reparameterized in the term of dispersion by letting $\phi = 1/\theta$. Then, the count y follows NB with mean = μ_s and variance = $\mu_s(1 + \phi\mu_s)$, where ϕ denotes the dispersion parameter. The farther ϕ falls above 0, the greater the overdispersion relative to Poisson variability. Clearly, when $\phi \to 0$, the NB distribution reduces to the usual standard Poisson distribution with parameter μ_s. In GLMs, the most convenient way to link the mean response μ of NB variable to a linear combination of the predictors X is the log link, as in Poisson loglinear models, for each feature i ($i = 1, 2, \ldots, r$), $log(\mu_s) = x_s^T\beta$, where x_s^T is $1 \times K$ row vector of indicator variables for the phenotypes, $s = 1, 2, \ldots, N$, K represents the number of phenotypes in the dataset and β is the corresponding $K \times 1$ column vector of unknown regression parameters (note: β here is different from the coefficient(s) β in the first stage. We still use the same symbol for the purpose of regression models). The covariates can be introduced into a regression model based on the NB distribution via the relationship

$$\log(\mu_s) = \sum_{j=1}^{K} x_{sj}\beta_{j-1} \quad (6)$$

In the NB model for mean $\mu_s = exp(x_s^T\beta)$, β and ϕ are estimated by maximizing the log-likelihood function:

$$\ell(\beta,\phi;Y) = \sum_{s=1}^{N}\left\{ \log\left(\frac{\Gamma(y_s+\phi^{-1})}{\Gamma(\phi^{-1})}\right) - log(y_s!) - (y_s+\phi^{-1})log(1+\phi\mu_s) + y_s log\phi + y_s x_s^T\beta \right\} \quad (7)$$

More details on regression models for NB responses can be found in Cameron and Trivedi (1998).

Hypothesis testing of model parameters in phenotype comparison for each feature

To test the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$, denote the maximum value of likelihood function by ℓ_0 under H_0, and ℓ_1 under H_1, which states that at least one coefficient $\beta_j \neq 0$. H_0 stands for no phenotype effect, i.e. the feature is not differentially abundant across different phenotypes. The *likelihood ratio* test statistic is:

$$-2\log(\ell_0/\ell_1) = -2[\log(\ell_0) - \log(\ell_1)] = -2(L_0 - L_1) \quad (8)$$

where L_0 and L_1 are the logarithms of maximum likelihood functions. Under H_0 this test statistic has an asymptotically chi-squared distribution with $K-1$ degrees of freedom.

Multiple test correction

A typical metagenomic dataset consists of several hundreds or thousands of features. After comparing multiple metagenomic groups using GLMs with the NB canonical logarithmic link function for simultaneous comparison, we used Benjamini–Hochberg's procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR) at significance level of 0.05.

3 SIMULATION STUDIES

Because of high similarity between RNA-Seq and metagenomic data, the statistical methods developed for RNA-Seq data in detecting differentially expressed genes may be applicable to the analysis of metagenomic data. For this reason we compared our method with two widely used statistical packages for RNA-Seq analysis, edgeR and DESeq, in addition to metagenomeSeq. As Metastats approach can only be used to

Table 1. The ranges of means of the NB distributions in four simulation settings

Setting	Minimum $\log_{10}$ (mean)	Maximum $\log_{10}$ (mean)
1 (Low)	0	1
2 (Intermediate)	1	2.5
3 (High)	2.5	5
4 (Combined)	0	5

Notes: Settings 1–4 reflect the count data of feature abundances with low means, intermediate means, high means and a combination means, respectively. Setting 4 most resembles to the nature of real metagenomic dataset.

compare two conditions/phenotypes, we also evaluated its performance in the following designs of two-condition or -phenotype comparison.

3.1 Experimental data

To make simulated data to reflect the nature of real metagenomic data we examined several types of real datasets from various environmental sources, including human gut, ocean, soil and fresh water (Supplementary Table S1), and obtained the means and variances of feature abundance in these studies. Interestingly, we observed strong linear relationships between the $\log_{10}$-transformed means of the feature abundances and the $\log_{10}$-transformed variances of abundances (Supplementary Figs S1–S4).

Experimental Design 1 **(two-condition comparison + fixed parameters for simulating data)**

We designed a metagenomic simulation study in which samples are drawn from two conditions. Because the sample size affects the performance of statistical methods, we designed metagenomic datasets with various sample sizes, including 10, 25 and 50 subjects drawn from each population. For each dataset, counts were generated using NB distributions, with different means (μ) and variances (σ^2). The means (μ) of the NB distributions were selected by random sampling from the ranges of the means for the abundances in four simulation settings (Table 1), and then the corresponding variances were computed from the following function:

$$\log_{10}(\sigma^2) = \beta_0 + \beta_1 * \log_{10}(\mu) \quad (9)$$

(In the first experiment let $\beta_0 = 0.6$ and $\beta_1 = 1.8$, which are from the observation of four real metagenomic datasets; details can be found in the supplementary file. In next two experiments, we will relax these two values). In each dataset, we simulated 1000 features for each sample of two conditions from NB distributions; 950 of them were generated from the same NB distribution, i.e. $\mu_1 = \mu_2$ with the corresponding variances computed by (9), and the rest 50 were generated from two different NB distributions, i.e. $a * \mu_1 = \mu_2$, where the parameter a (i.e. multiplier) is selected from the set of 1.5, 2.5, 5, 7.5 and 10. To prevent bias arising from a specific partition, we simulated the datasets 100 times for each sample size. The performance of four methods were compared using the 'area under the curve' (AUC) metric of a receiver operator curve (ROC), and the true-positive rate (tpr, i.e. power) were calculated at each level of FDR.

Experimental Design 2 **(two-condition comparison + varied parameters for simulating data)**

Different from the first experimental design where the values of β_0 and β_1 are fixed, the second experiment allows these two parameters to vary. They were determined by random sampling from the ranges of [0.1, 1] and [1.5, 2], respectively. These ranges of the estimates for β_0 and β_1 were obtained from observing real metagenomic data (details in Supplementary). As the setting 4 resembles most to the nature of real metagenomic dataset, in the second experiment we flexed the β_0 and β_1 on this setting. Similar to the first experiment, we simulated 1000 features for each sample of the two conditions from NB distributions: 950 of them were generated from the same NB distribution, and the rest were from two different NB distributions.

Experimental Design 3 **(three-condition comparison + varied parameters for simulating data)**

In this experiment the samples were drawn from three conditions. The parameter settings for β_0 and β_1 are as same as the Experimental Design 2. For each sample under different conditions we simulated 1000 features from NB distributions: 950 of them were generated from the same NB distribution, and the rest 50 features were from different NB distributions. That is, at least two NB distributions (representing two conditions) of three distributions are different for each of these 50 features.

3.2 Simulation results

Results from Experimental Designs 1 and 2

ROC curve is usually used in measuring signal detection. It is created by plotting the true-positive rate versus the false-positive rate. AUC shows an overall performance of detection methods. The higher the AUC value, the better the method is. Figure 1 displays the AUC results for four methods with different sample sizes (10, 25 and 50) under four simulation settings. AUC values generally increase when the sample size increases; AUC values are greater for higher mean setting. The proposed approach outperforms the other methods in the Setting 2 for small sample size (n = 10) and in Setting 1 for large sample size (n = 50) and is well comparable with other methods in the rest of the settings.

In addition to the AUC, which shows an overall performance of the methods, we also compare our method with other methods in terms of power in detecting truly differentially abundant features, while the FDR is controlled at different levels. Figure 2 shows the power for sample size of 10, 25 and 50 in each simulation setting in Experimental Design 1. For the Settings 2–4 our proposed approach either outperforms other methods or is well comparable with other methods. In the Setting 1, the new method surpasses other methods for sample size of 50, and is comparable with metagenomeSeq for sample size of 25 while much powerful than the rest. Interestingly, for sample size of 10, metagenomeSeq shows much higher power than the new method. We examined the true (i.e. realized) FDR at adjusted *P*-value of 0.05. The boxplots of true FDR across 100 replications for this experiment show that the FDR cannot be

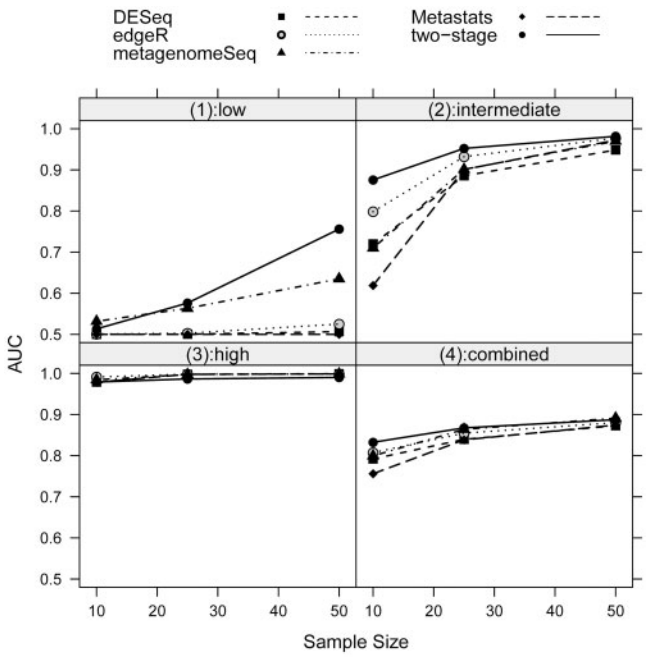

Fig. 1. The AUC results for sample size of 10, 25 and 50 in each simulation setting in the experimental design 1. (1–4) show the AUC results for four settings, i.e. low means, intermediate means, high means and combination of means, respectively

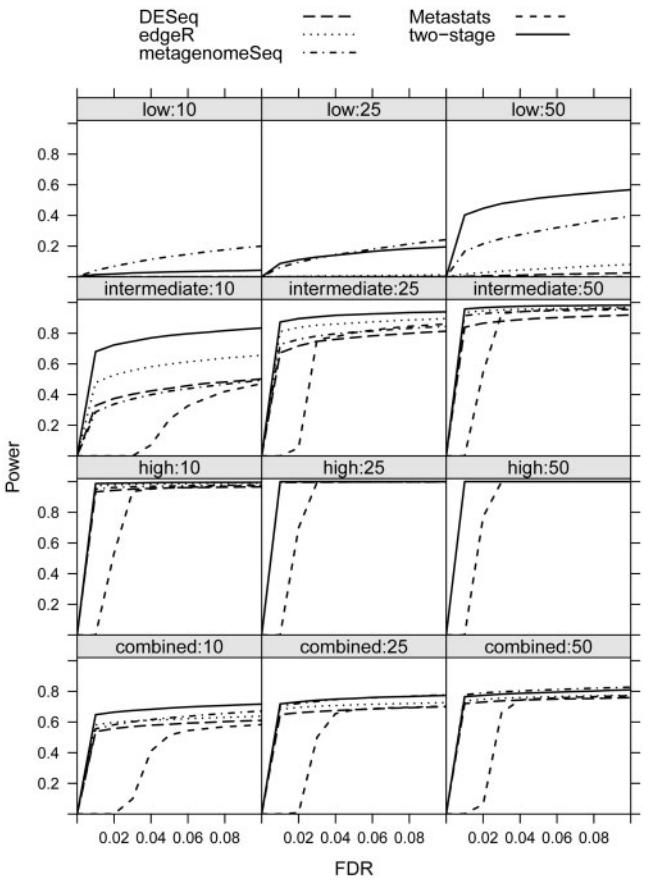

Fig. 2. The power in detection of the true differentially abundant features for four methods at various levels of FDR for sample size of 10, 25 and 50. (1–4) show the power for four settings in the first experiment, i.e. low means, intermediate means, high means and combination of means, respectively

controlled well for metagenomeSeq for any sample size in Setting 1 and 4 (Supplementary Fig. S5). The same conclusion can be obtained for true FDR plots at adjusted *P*-value of 0.01 (Supplementary Fig. S6).

The AUC plots and power plots for Experimental Design 2 can be found in the Supplementary Figures S7 and S8. Both types of plots demonstrate that the new method outperforms others, in particular, when the sample size is small. The true FDR plots (Supplementary Fig. S5 and S6) indicate that FDR is not controlled by metagenomeSeq for any sample size in the Experiment 2.

Results from Experimental Design 3

Figure 3 displays the AUC results, and Figure 4 shows the power detection of truly differentially abundant features obtained from each method for sample size of 10, 25 and 50 in the simulation setting in the Experimental Design 3. The AUC results show that the proposed method and edgeR outperform other methods in situations with sample size of 10 and 25 and are comparable with DESeq and Metastats in a situation with large sample size. The proposed approach has highly similar performance with edgeR when both of them are compared in terms of AUC as shown in Figure 3. However, our proposed method outperforms edgeR and other methods in terms of the power in detecting the true differentially abundant features as shown in Figure 4, in particular, when the sample size is small. For FDR, metagenomeSeq is the only method that is a little above the reference horizontal line (0.05, Supplementary Fig. S5).

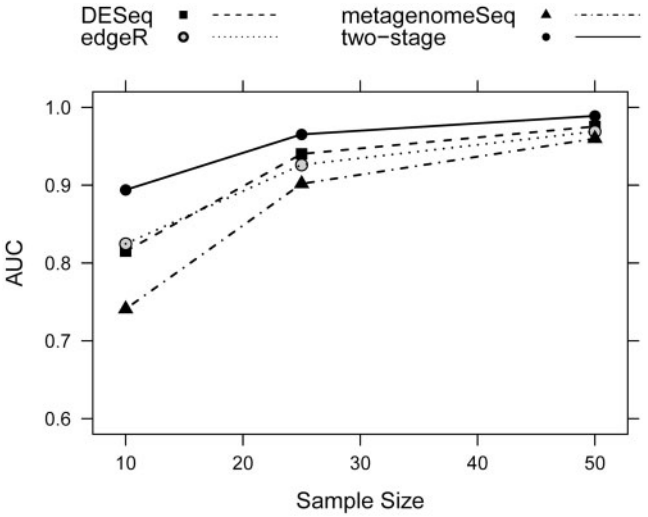

Fig. 3. The AUC results for sample size of 10, 25 and 50 in the simulation setting in the Experimental Design 3

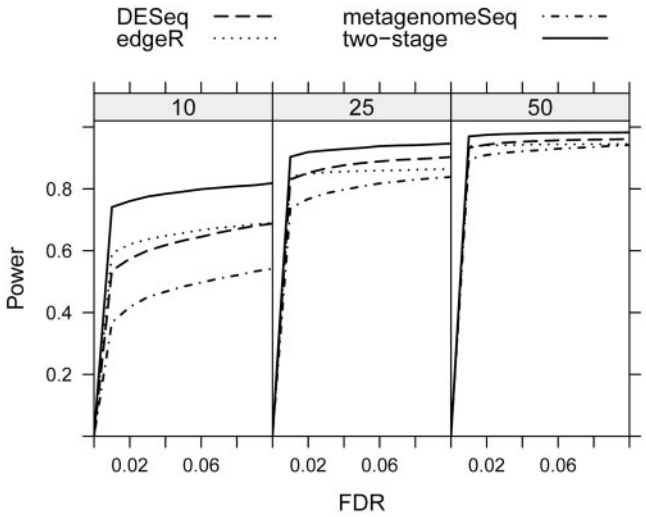

Fig. 4. The power in detection of the true differentially abundant features obtained from each method for sample size of 10, 25 and 50 in the Experimental Design 3

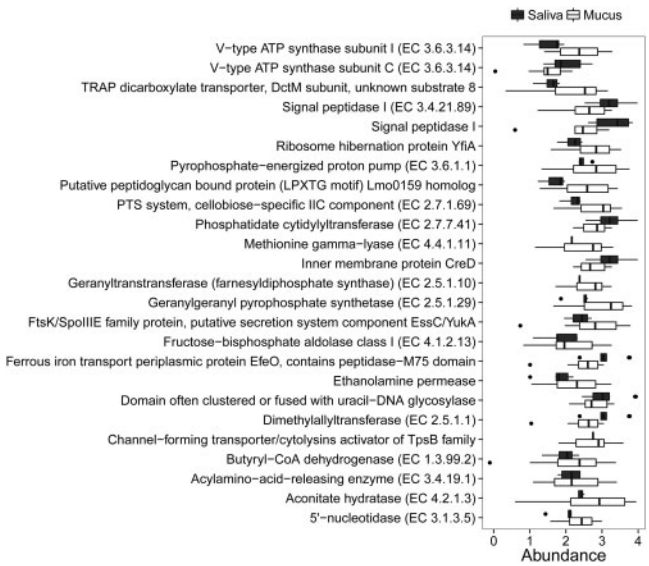

Fig. 5. Differentially abundant functions (in $\log_{10}$ scale) between human mucus and human saliva individuals

Table 2. Comparison of computational time (in second) for five methods on one simulated dataset (Setting 4 of Experiment 1) under various sample sizes

Sample size	DESeq	edgeR	metagenomeSeq	Metastats	Two-stage
N = 10	325.66	1.15	4.92	220.54	6.49
N = 25	593.46	2.40	5.71	532.24	7.30
N = 50	787.73	4.21	7.22	535.83	10.32

Computational time

A comparison of computational time for five methods on one simulation dataset of Setting 4 is shown in Table 2. The simulation was done on a PC with 2.33 GHz and 4.00 GB RAM. Two-stage, edgeR and metagenomeSeq are comparable while DeSeq and Metastats take 30–300 times longer. The computational time for other settings are similar. Note the results shown in Figure 1–4 are for 100 repetitions, while the Table 2 shows the time for one repetition.

4. Real data analysis

Human mucus versus saliva data

We performed our proposed method on metagenomic shotgun sequence data in the HMP project (Qin *et al*., 2010) focusing on the functions of microbes in human health and disease through the characterization of microbial communities for two human body sites: nasal mucus and oral saliva. Of 42 samples, 30 samples are obtained from human nasal mucus microbial metagenomes and 12 samples from human oral saliva samples. The dataset was downloaded from MG-RAST.

Differentially functional abundances between human nasal mucus and human oral saliva were identified with multiple comparison correction of FDR < 0.05. Figure 5 shows the top 25 most significant differentially abundant functions. Five of them get involved in a biological process of phosphate metabolism, and their abundances are more presented in microbial metagenomes of cystic fibrosis (CF) lung patients compared with microbial metagenomes of healthy human saliva individuals. These functions are *Pyrophosphate-energized proton pump (EC 3.6.1.1)*, *Geranyltranstransferase (farnesyldiphosphate synthase) (EC 2.5.1.10)*, *Geranylgeranyl pyrophosphate synthetase (EC 2.5.1.29)*, *Fructose-bisphosphate aldolase class I (EC 4.1.2.13)* and *Maltose-6′-phosphate glucosidase (EC 3.2.1.122)*. Willner *et al*. (2009) conducted the first metagenomic study of DNA viral communities in the airways of CF diseased and non-diseased individuals and discovered that *Guanosine-5′-triphosphate, 3′-diphosphate pyrophosphatase* are over-representation in CF diseased compared with non-diseased individuals. Several studies, including Jain *et al*. (2006) and Raskin *et al*. (2007), discovered that these enzymes are linked to bacterial stringent response, bacterial virulence, antibiotic resistance, biofilm formation, quorum sensing and phage induction in a variety of bacteria. These findings imply that a unique metagenomic environment of the CF airway might contribute to functional adaptations, resulting in shifts in metabolic profiles (Willner *et al*., 2009).

Moreover, we found that *Putative peptidoglycan bound protein (LPXTG motif) Lmo0159 homolog* is enriched in mucus but rare in saliva metagenomes. This finding is correspondent to the discovery of Quinn *et al*. (2014), which conducted an experiment to assess how CF lung microbes respond to the biochemistry of the lung environment by identifying pathways, obtained from KEGG classification hierarchy, whose presence enriched in microbial metagenomes of CF lung patients compared with healthy human saliva microbial metagenomes from the HMP. Quinn *et al*. (2014) reported that peptidoglycan biosynthesis pathway is enriched in human mucus metagenomes of CF lung patients,

but rare in healthy human saliva individuals. Furthermore, of the significant differentially abundant functions, we discovered that three functions, including *Glutamate formyltransferase, Formiminoglutamase (EC 3.5.3.8)* and *Aminobenzoyl-glutamate transport protein*, are involved in glutamate protein and are enriched in human mucus. Our finding is also consistent to the findings discovered by Quinn *et al.* (2014), that D-glutamine and D-glutamate metabolism pathways are enriched in human mucus of CF lung patients compared with healthy human saliva. The results suggest that enrichment of those functions in human mucus of CF lung patients compared with healthy human saliva individuals may be a contributor to CF disease.

Human gut data

We applied our proposed method on human gut metagenomic data from 124 unrelated Danish and Spanish individuals in the Meta-HIT project (Qin *et al.*, 2010) focusing on two human diseases, obesity and inflammatory bowel disease (IBD). The occurrence of obesity patients with IBD has become increasingly prevalent over the past two decades (Boutros and Maron, 2011). The DNA sequences were aligned to the MetaHIT gene catalogue of 3.3 million genes to get the abundance of genes. The genes were annotated to the NCBI non-redundant Clusters of Orthologous Groups (COGs) database, and this information was used to transform gene abundance to COG abundances. Of the 124 individuals, 82 were labeled as lean [body mass index (BMI) < 30] and 42 were labeled as obese (BMI ≥ 30). Moreover, 3 of 42 obese people were diagnosed with IBD and 22 of 82 leans with IBD. Thus, we have four phenotypes or groups for comparison. Differences based on our two-stage method with multiple comparison correction of FDR < 0.05 are observed among the four groups in COG functional terms. Figure 6 displays the top 25 most significant functions whose abundance differs among the four groups.

We found that two *cytochrome c biogenesis* involved functions *Cytochrome c-type biogenesis factor* and *Cytochrome c-type biogenesis protein CcmE* in obese only group are significantly differentially abundant comparing with healthy group or IBD and obese group. And their abundances are marginally significantly different from IBD-only group. This may imply that even though, in general, obesity increases the risk of IBD, obesity caused by lack of *cytochrome c* may not increase the risk of IBD. Hence, the alteration of *cytochrome c* can potentially be used as a biomarker to help stratify obese patient by the risk of developing IBD.

The top significant COGs (adjusted overall P-value < 0.01), along with the adjusted P-values for each pairwise comparison, are given in the Supplementary Table. By pairwise comparison in this table, we also found the count of *asparaginase* in IBD and obese group is significantly different from IBD-only group, obese-only group and healthy group, respectively, whereas, the other pairwise comparisons for *asparaginase* did not yield any significant result. This suggested that *asparaginase* might only contribute to IBD when the patient is obese. In 2013, Ehsanipour *et al.* (2013) showed that obesity impaired L-asparaginase treatment due to the fact that *adipocytes* work in conjunction with other cells of the leukemia microenvironment. For IBD patients, it is possible that *adipocytes* play a role in the interaction between IBD cells and *asparaginase*, which explains why the count of asparaginase only differs in the IBD and obese group.

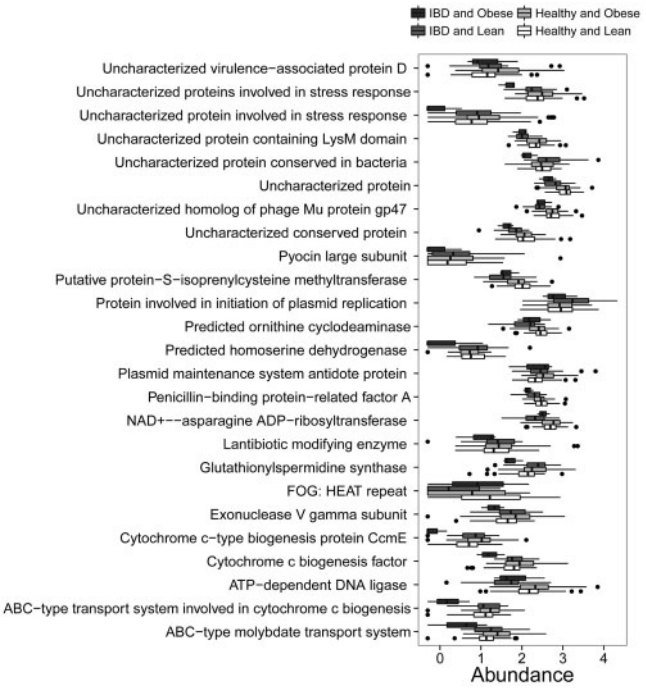

Fig. 6. Differentially abundant functions (in $\log_{10}$ scale) among the four groups

5 DISCUSSION

Currently, there has been an increasing interest in metagenomic projects with various applications. One typical aim is to assess whether and how two or more microbial communities differ. Comparing microbial genetic contents on the basis of functional features (e.g. pathways, subsystems, functional roles) obtained from different microbial communities with different phenotypes (e.g. diseased and healthy, or different treatments) enables us to identify the gnomic contents of microbes contributing to human health and disease, which can in turn lead us to understand how the microbes affect human health.

We proposed a two-stage statistical procedure for sequentially selecting informative functional features and detecting differentially abundant functional features between two or more microbial communities/conditions. The proposed method accounts for the specific characteristics of metagenomic data, which are high-dimensional complex datasets consisting of a large proportion of zeros, non-negative counts with skewed distribution and a large number of features, but limited number of samples. From the results of various simulations, we showed that our proposed method more effectively selects the informative functional features and therefore more efficiently detects the differentially abundant functional features between metagenomic datasets. Owing to the existence of large proportion of zeros in metagenomic data, we also fitted the Zero Inflated Negative Binomial

(ZINB) on the filtered data through elastic net for the Experiment 1. Comparing the results from NB and ZINB methods, NB approach exceeds the ZINB fitting for most of the cases according to the AUC plots and power plots (shown in the Supplementary File, Supplementary Figs S9 and S10); otherwise these two methods are comparable. However, the computational time for ZINB is ~200–300 times longer than for NB fitting due to more parameters in the ZINB models.

We also applied the proposed method on two real metagenomic datasets related to two human diseases. One of them is related to obesity and IBD, and the other one is related to CF lung disease. In the gut data, there are four phenotypes/groups owing to the combination of the two diseases. Our method is directly applied to this multiple-group comparison and our findings are consistent with previous reports. Compared with other existing methods on metagenomic studies, the proposed two-stage method is more powerful and flexible.

Funding: This work was supported by National Science Foundation [DMS-1043080 and DMS-1222592 to L.A. and H.J.], and partially supported by National Institutes of Health [P30 ES006694 to L.A.] and by The Cecil Miller Endowment at University of Arizona Foundation (to N.P.)

Conflict of interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Boutros,M. and Maron,D. (2011) Inflammatory bowel disease in the obese patient. *Clin. Colon Rectal. Surg.*, **24**, 244–252.

Cameron,A. and Trivedi,P. (1998) *Regression Analysis of Count Data*. First Edition. Econometric Society Monograph No. 30, Cambridge University Press.

Ehsanipour,E.A. *et al.* (2013) Adipocytes cause leukemia cell resistance to *L-Asparaginase* via release of glutamine. *Cancer Res.*, **73**, 2998–3006.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw. Jan.*, **33**, 1–22.

Gilbert,J.A. *et al.* (2011) The future of microbial metagenomics (or is ignorance bliss?). *ISME J.*, **5**, 777–779.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. 2nd edn. Springer-Verlag, New York, NY.

Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, REVIEWS0003.

Huson,D. *et al.* (2009) Methods for comparative metagenomics. *BMC Bioinformatics*, **10** (**Suppl. 1**), S12.

Huson,D. *et al.* (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, **21**, 1552–1560.

Jain,V. *et al.* (2006) ppGpp: stringent response and survival. *J. Microbiol.*, **44**, 1–10.

Kristiansson,E. *et al.* (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics*, **25**, 2737–2738.

Kunin,V. *et al.* (2008) A bioinformatics's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, **72**, 557.

Paulson,J. *et al.* (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, **10**, 1200–1202.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Quinn,R.A. *et al.* (2014) Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *mBio*, **5**, e00956–13.

Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

Robinson,M. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

Robinson,M. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodriguez-Brito,B. *et al.* (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**, 162.

Raskin,D.M. *et al.* (2007) Regulation of the stringent response is the essential function of the conserved bacterial G protein CgtA in Vibrio cholerae. *Proc. Natl Acad. Sci. USA*, **104**, 4636–4641.

Schloss,P. and Handelsman,J. (2006) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl. Environ. Microbiol.*, **72**, 6773–6779.

Turnbaugh,P. *et al.* (2007) The human microbiome project. *Nature*, **449**, 804–810.

Venables,W. and Ripley,B. (2002) *Modern Applied Statistics with S*. 4th edn. Springer-Verlag, New York, NY.

White,J. *et al.* (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352.

Willner,D. *et al.* (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*, **4**, e7370.

Wooley,J. and Ye,Y. (2010) Metagenomics: facts and artifacts, and computational challenges. *J. Comp. Sci. Tech.*, **25**, 71–81.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301–320.