Bioinformatics, 31(10), 2015, 1607–1613
doi: 10.1093/bioinformatics/btu855
Advance Access Publication Date: 6 January 2015
Original Paper

Systems biology

Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data

Qunyuan Zhang[1,*], Haley Abel[1], Alan Wells[1], Petra Lenzini[1], Felicia Gomez[1], Michael A. Province[1], Alan A. Templeton[1,2], George M. Weinstock[3], Nita H. Salzman[4] and Ingrid B. Borecki[1]

[1]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA, [2]Department of Biology, Washington University, St. Louis, MO, USA, [3]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA and [4] Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 26, 2014; revised on November 27, 2014; accepted on December 23, 2014

Abstract

Motivation: Establishment of a statistical association between microbiome features and clinical outcomes is of growing interest because of the potential for yielding insights into biological mechanisms and pathogenesis. Extracting microbiome features that are relevant for a disease is challenging and existing variable selection methods are limited due to large number of risk factor variables from microbiome sequence data and their complex biological structure.

Results: We propose a tree-based scanning method, Selection of Models for the Analysis of Risk factor Trees (referred to as SMART-scan), for identifying taxonomic groups that are associated with a disease or trait. SMART-scan is a model selection technique that uses a predefined taxonomy to organize the large pool of possible predictors into optimized groups, and hierarchically searches and determines variable groups for association test. We investigate the statistical properties of SMART-scan through simulations, in comparison to a regular single-variable analysis and three commonly-used variable selection methods, stepwise regression, least absolute shrinkage and selection operator (LASSO) and classification and regression tree (CART). When there are taxonomic group effects in the data, SMART-scan can significantly increase power by using bacterial taxonomic information to split large numbers of variables into groups. Through an application to microbiome data from a vervet monkey diet experiment, we demonstrate that SMART-scan can identify important phenotype-associated taxonomic features missed by single-variable analysis, stepwise regression, LASSO and CART.

Availability and implementation: The SMART-scan approach is implemented in R and is available at https://dsgweb.wustl.edu/qunyuan/software/smartscan/

Contact: qunyuan@wustl.edu

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

1 Introduction

In recent years, the problem inherent in characterizing microbiome communities is of rapidly growing interest in considering DNA sequences extracted from samples of these communities. Application of this method of interrogation to large samples of individuals in population-based studies has become possible due to advances of technology and the significant reduction in sequencing costs, and investigators are increasingly considering the role of microbial composition in human health and disease. Microbiome research is an important, rapidly growing field. One of its goals is the identification of features (such as diversity or composition) of microbial communities and their association with host phenotypes. Typically, the 16S rRNA gene amplicons from samples are sequenced and classified into taxonomic trees, based on phylogenetic relationships, through reference alignment and/or unsupervised learning, and then the association between tree structures and host phenotypes is investigated. In such studies, the statistical analysis is challenging because of the large number of tree variables and hierarchical, nested tree structures.

A variety of statistical methods have been proposed for detecting the association between microbiome features and host phenotypes. Microbiome data commonly used for these analyses are sequence counts (usually standardized to relative abundances or dichotomized to presence or absence) of operational taxonomic units (OTUs). OTUs represent the lowest taxonomic level into which a sequence can be assigned, that is typically used in downstream analysis. An OTU can be defined as a phylogenetic unit (such as bacterial species or genus) or a set of similar DNA sequences learned by data clustering (Wooley *et al.*, 2010). For many analyses, a phylogeny relating the OTUs is needed. The process of determining OTU counts, referred to as OTU picking, can be reference-based, i.e. sequence reads are mapped to a reference database of known 16S rRNA gene sequences, using, for instance, the Ribosomal Database Project (RDP) classifier (DeSantis *et al.*, 2006; Pruesse *et al.*, 2007; Wang *et al.*, 2007). In this case, the OTUs have a presumed known phylogeny and taxonomy established externally and documented in the reference database; for many of the common species, these relationships are well known. In *de novo* OTU picking, reads are clustered based on sequence similarity and OTUs are defined based on a maximum dissimilarity threshold (Caporaso *et al.*, 2010; Schloss *et al.*, 2009). Here, phylogeny can be inferred either *de novo* (e.g. by constructing a neighbor-joining tree) or by mapping representative reads to a reference database (Felsenstein, 2005; Ludwig *et al.*, 2004). In either case, once a phylogenetic tree has been constructed, the information comprising the tree can be leveraged in downstream analyses.

Downstream statistical analysis methods can be categorized into several broad approaches: simple univariate methods, tests of diversity, multivariate methods, and model-selection methods. Univariate approaches generally ignore the phylogenetic tree structure and multivariate and correlated nature of the OTU count data; these include simple one-taxon-at-a-time tests of association between phenotype and OTU abundance (or presence/absence), and are likely to be powerful when one or a few taxa have strong effects on host phenotype (White *et al.*, 2009). Another relatively simple approach is to associate microbiome diversity with host phenotype. Numerous measures of diversity exist and provide different ways to quantify the 'richness' or 'evenness' of OTUs within a phenotype (Hill *et al.*, 2003; Holmes *et al.*, 2012).

Multivariate methods that assess the association of an entire vector of multiple OTUs have been widely used. As OTU abundances are nonnegative and not distributed as a multivariate normal, standard multivariate methods (e.g. Hotelling T^2 test and multivariate analysis of variance (MANOVA)) are generally inappropriate. Methods have been developed to model the OTU abundances as drawn from a Dirichlet-multinomial, or mixture of Dirichlet multinomials (Holmes *et al.*, 2012; La Rosa *et al.*, 2012). Nonparametric multivariate methods for hypothesis testing are commonly used as well. In general, these are defined based on a distance measure of dissimilarity between each pair of samples, with the samples characterized by a vector of OTU counts. The distance measure can be tree unaware (e.g. Euclidean distance or Bray–Curtis dissimilarity) or tree aware. Tree aware measures can take into account the structure of the phylogenetic or taxonomic tree. For example, UniFrac is a tree-aware method that calculates distances between pairs of microbiome samples based on the fraction of the total branch length leading uniquely to one or the other community (Lozupone and Knight, 2005). Based on pairwise distances, various nonparametric multivariate tests can be conducted, such as the Mantel test and nonparametric MANOVA (Anderson, 2001; Chen *et al.*, 2012; Mantel, 1967).

A final category of methods for analyzing association between OTU abundances and host phenotypes comprises regression-based variable-selection methods. Due to the large number of parameters (usually coupled with relatively small sample size) in microbiome data analysis, l1-regularized regression, in particular the least absolute shrinkage and selection operator (LASSO), has been considered as an appropriate choice (Sun *et al.*, 2010; Tibshirani, 1996).

In summary, these four categories of methods were developed to address two types of questions. While diversity tests and multivariate methods are usually used to determine whether the overall feature (diversity or composition) of an entire tree of OTUs is associated with a phenotype (i.e. overall association), univariate tests and model-selection methods identify which specific OTUs are associated (i.e. specific association). In practice, a specific association is particularly interesting and important, because it can provide biological insight into the mechanism of the disease and help define a potential target for modification. However, existing methods for detecting specific association (such as univariate or multivariate regression, LASSO, etc.) have significant limitations. Most of these methods do not take into account the taxonomic or phylogenetic relationship among OTUs. In these methods, individual OTUs are modeled as separate variables independent of each other, although shared evolutionary ancestry means that OTUs cannot be treated as statistically independent groups (Felsenstein, 1985, 2008). Another issue is that the microbiome data can be classified across multiple levels and an association can be driven by an enrichment of an entire group of OTUs at higher taxonomic levels such as bacterial families, orders or even phyla (Ahn *et al.*, 2013; Dey *et al.*, 2013; Garcia *et al.*, 2013; Segata *et al.*, 2012). In this case, only modeling individual OTUs at the lowest taxonomic level may significantly lose power because of weaker effects of individual OTUs. Although existing univariate tests and variable selection methods (even tree-based multivariate methods such as UniFrac) can be applied to different levels or subsets of the data to identify specific association, conducting such analyses of all possible levels and subsets of a large taxonomic tree creates an intractable multiple comparison problem, therefore, strategies leveraging additional biological information to reduce the model space and optimize the search procedure are desirable.

To overcome these issues, we have developed a novel tree-based method, Selection of Models for the Analysis of Risk-factor Trees (SMART-scan), which can leverage phylogenetic knowledge to mine large sets of tree risk factors and is particularly appropriate and powerful for detecting the OTUs associated with host phenotypes and grouped by similar effects. Unlike those widely-used multivariate methods (e.g. UniFrac) for overall association analysis using pairwise distances between subjects or hierarchical tree of subjects, SMART-scan is designed for specific association analysis using the tree of risk factors (i.e. OTUs), not the tree of subjects. Essentially, SMART-scan is a variable grouping technique that incorporates the known structure of the variables into models. It adopts a tree scanning strategy that efficiently performs an iterative search for risk factors (or groups of factors) within a structured tree. This strategy was previously introduced and used for detecting phenotypic associations with haplotype trees at candidate loci (Templeton *et al.*, 2005). In this article, we present the theory and algorithm of SMART-scan, and, in comparison to other four existing methods, we investigate its statistical properties using simulated data and demonstrate its application using microbiome data from vervet monkeys.

2 Methods

2.1 Assumptions and models

SMART-scan is based on a biological assumption that multiple, closely-related OTUs in a tree may function in a similar manner as a group because of shared evolutionary ancestry. Evidence supporting this assumption can be found in many studies, e.g. an entire bacterial family of *Streptococcaceae* has been reported to have an effect on the CD68 expression (Garcia, 2013); a significantly lower relative abundance of all members of the phylum *Firmicutes* has been observed in the samples from Crohn's patients with recurrence relative to those who stayed in remission (Dey, 2013); colorectal cancer patients tended to have enrichment of phylum *Bacteroidetes* and depletion of *Firmicutes* (Ahn, 2013). Such observations suggest that the effects of OTUs on some phenotypes may come from one or more groups. The goal of SMART-scan is to identify such group(s) under the guide of a phylogenetic tree. When there is no group effect (i.e. individual OTUs have very different effects), SMART-scan is still helpful in model building by compressing large number of null OTUs into a single or a few noise groups.

For the convenience of discussion, let **Y** denote the phenotype data of N samples, $\mathbf{X} = (x_1, x_2, x_3, \ldots, x_m)$ the microbiome data of m OTUs observed by sequencing the samples (where x_i is a vector of N observations for the i^{th} OTU), and $\mathbf{T}_m$ the taxonomic or phylogenetic tree structure of the OTUs. Here **Y** can be qualitative or quantitative; **X** can be any types of transformed sequence data, such as relative abundance, logarized abundance, presence or absence of reads, etc. We choose to use relative abundance in this paper. It is defined as the percentage of the mapped sequence reads of an OTU relative to the total mapped reads from a sample and has been widely used in microbiome studies to measure the OTU composition. $\mathbf{T}_m$ represents the biological knowledge (i.e. similarity or phylogenetic information) on the m OTUs. In microbiome analysis, a tree of OTUs can be obtained either by mapping OTUs to reference sequences of a known taxonomy or through an unsupervised hierarchical clustering. SMART-scan is a variable grouping procedure in which the structure information from $\mathbf{T}_m$ is used for building models. The basic idea of SMART-scan is to find the optimal OTU groups of **X** with the aid of the similarity or phylogenetic information from $\mathbf{T}_m$, and then test the association between each group and phenotype.

To describe the SMART-scan model, we use eight OTUs (i.e. $m = 8$) as a simple example. When $m = 8$, $\mathbf{X} = (x_1, x_2, x_3, \ldots, x_8)$, a regular multivariate regression model for **Y** can be defined by Equation (1) with parameters intercept (α), regression coefficients (β_1 to β_8) and random error (e).

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_8 x_8 + e \tag{1}$$

In real microbiome data, m could be large (usually hundreds, even thousands) and such a model may significantly lose power, due to the small number of degrees of freedom (especially when the sample size is small) as well as the large number of tests. Another major limitation is that the abundance of some OTUs is very low and distributed very sparsely among host samples. The power to detect the association of these OTUs could be very low when they are individually modeled in a regular multivariate regression model. To improve the power, instead of modeling variables individually, SMART-scan uses a group model, e.g.:

$$Y = a + b_1 Z_1 + b_2 Z_2 + b_3 Z_3 + e \tag{2}$$

In this model, eight OTUs are divided into three groups and summed into three new variables Z_1, Z_2 and Z_3 (e.g., $Z_1 = x_1 + x_2$, $Z_2 = x_3 + x_4$ and $Z_3 = x_5 + x_6 + x_7 + x_8$). Accordingly, the data are modeled by Equation (2) with parameters intercept (a), regression coefficients (b_1, b_2 and b_3) and random error (e). Such a group model is expected to be more powerful than the regular multivariate model when the OTUs within each group share the same or similar effects (e.g. both x_1 and x_2 have large effects, x_3 and x_4 small effects, x_5, x_6, x_7 and x_8 no effects), because it can reduce the total number of parameters to be estimated and tested in the model, increase the degrees of freedom, alleviate the multiple testing issue, and make the data less sparse via summarization, thereby improving the power of the association test.

2.2 Algorithm

A challenge of applying such a group model to microbiome data is that an actual dataset from microbiome experiments usually contains hundreds of OTU variables; therefore, the possible ways of grouping these variables are too many to be evaluated exhaustively. To deal with this issue, we propose the SMART-scan procedure which uses the OTU tree to optimize the grouping of variables in a more efficient and biologically meaningful way. The logic of SMART-scan is that an OTU tree can provide prior knowledge on the similarities and phylogenetic effects of shared ancestry between the OTU variables. Such knowledge can be used to exclude many unlikely group models, thus limiting the possible models to a significantly smaller space compatible with the tree, and making the model search feasible even for a large number of OTUs. To realize this with an acceptable computational burden, SMART-scan determines the group model through an iterative, nested search procedure using the Akaike information criterion (AIC), as described below and illustrated in Figure 1.

Given a tree of OTUs, SMART-scan first performs an initial scan to identify the best group model that contains only two groups of OTU variables. The best model is defined as the model that produces the smallest AIC. It is identified through comparing the AICs of all possible two-group models generated by cutting the tree at all possible branches (Fig. 1a). For each possible cut, the tree is split into two parts (i.e. two subtrees), resulting in a two-group model, $Y = a + b_1 Z_1 + b_2 Z_2 + e$, where Z_1 and Z_2 are the sums of the variables in the two subtrees, respectively. After the initial scan obtaining two subtrees (denoted by T_1 and T_2) and the smallest AIC (denoted by AIC_0) for the best cut, further and similar scans are performed within T_1 and T_2, separately, to obtain new models and

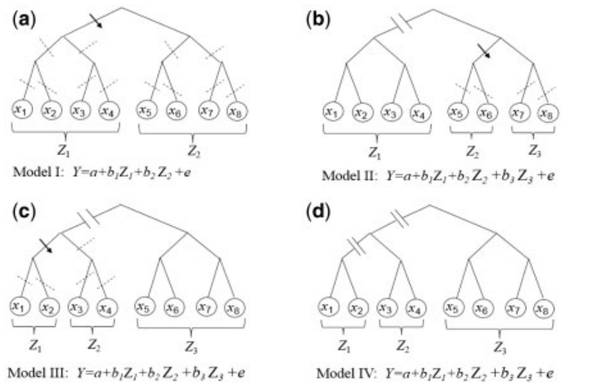
(a)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Z_1 Z_2
Model I: $Y=a+b_1Z_1+b_2Z_2+e$
(b)
Z_1 Z_2 Z_3
Model II: $Y=a+b_1Z_1+b_2Z_2+b_3Z_3+e$
(c)
Z_1 Z_2 Z_3
Model III: $Y=a+b_1Z_1+b_2Z_2+b_3Z_3+e$
(d)
Z_1 Z_2 Z_3
Model IV: $Y=a+b_1Z_1+b_2Z_2+b_3Z_3+e$

Fig. 1. Illustration of the SMART-scan algorithm. This figure illustrates the SMART-scan algorithm, using a small tree with eight OTUs ($x_1 - x_8$) as an example. It starts from (**a**) an initial scan that compares all possible cuts (denoted by dash lines) and identifies the best one (denoted by an arrow) producing the model (model I) with the smallest AIC and splitting the tree in to two sub-trees, followed by the second-round scans within (**b**) the right-side sub-tree and (**c**) the left-side sub-tree. Since the best model (model II) generated by the right-side sub-tree scan does not produce better (i.e. smaller) AIC than the previous model (model I), the cut is rejected. On the contrary, the best cut in the left-side sub-tree produces a model (model III) with better AIC and therefore is accepted. This procedure is iteratively repeated until no cuts with improved AIC are found, resulting in (**d**) a final model (model IV) which is produced by all accepted best cuts (denoted by double lines). In all models, Z_i is the sum over a set of *x* variables

AICs (Fig. 1b, c). If a new AIC generated by the second-round best cut is smaller than AIC$_0$, the corresponding cut of subtrees and the resultant model (with more groups of variables) will be accepted, and AIC$_0$ will be updated by the new model. This procedure is repeated iteratively until no smaller AICs for any new cuts are observed. Finally, the sums of variables within the groups determined by all the accepted cuts are treated as new variables and fitted into the final group model (Fig. 1d). The significance of each group is evaluated using the final model and, through permutations of **Y**, *P* values of groups can be corrected for multiple testing and adjusted for the inflation caused by a nonrandom grouping of variables (Supplementary Fig. S1). In our analysis, we calculated the adjusted *P* value as the percent of $P_{\max} \leq$ unadjusted *P* value, where $P_{\max}$ is a set of *n* maximum *P* values obtained from *n* permutations (each permutation produces one maximum *P* value).

Through the procedure above, SMART-scan systematically and iteratively searches the optimal groups of variables with the guide of the tree, which not only avoids exhaustive evaluation and thus substantially reduces the computational time, but also is expected to produce more interpretable results because of the use of biological information from a tree.

2.3 Methods for comparison

We compared SMART-scan with a regular and simple single-variable analysis (SVA) and three commonly-used variable selection methods: stepwise regression, LASSO and CART, using both the simulated and real data. Since dependent variables in our simulated and real data are both binary, we chose a logistic regression of each OTU on a binary trait for the SVA analysis. The stepwise analysis was performed using the glm() and step() functions in R 2.15.1. LASSO was performed using the R glmnet package (version 1.9.5), with the value of the tuning parameter λ chosen by ten-fold cross-validation. CART was performed using the R rpart package (version 4.1.8) with 100 cross validations.

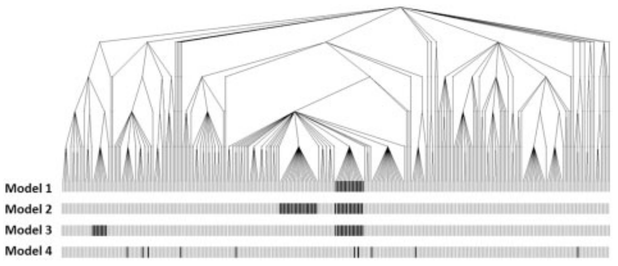

Fig. 2. A real microbiome tree and three different association models. The upper part of the figure shows the taxonomic tree of the 287 genera found in a vervet monkey microbiome diet experiment. The four rows of bars in the lower area represent the postulated effect of individual genera in four cases we simulated. The genera labelled in solid black are modelled as predictor variables in models 1–4, respectively. Model 1 (monophyletic model) contains 16 genera from a single bacterial family (*Ruminococcaceae*), model 2 (paraphyletic model) 25 genera from two close families (*Ruminococcaceae* and *Lachnospiraceae*), model 3 (polyphyletic model) 37 genera from two distant families (*Ruminococcaceae* and *Coriobacteriaceae*), and model 4 (non-phyletic model) 10 genera drawn from 10 random locations of the tree

3 Results

3.1 Simulation study

We simulated microbiome data based on a real phylogenetic tree with 287 bacterial genera, 156 families, 72 orders, 40 classes and 18 phyla (Fig. 2). Since the Dirichlet-multinomial distribution has been shown to be appropriate in modeling microbiome data (Holmes *et al.*, 2012; La Rosa *et al.*, 2012), we first estimated a Dirichlet-multinomial distribution using a set of real microbiome sequence count data from vervet monkey samples (see Section 3.2); then we simulated microbiome data based on the estimated Dirichlet-multinomial distribution. This allows us to simulate microbiome data with more realistic variation. All replications have the same OTU number of 287, the same tree structure (Fig. 2), and the same Dirichlet-multinomial distribution of OTU counts that is learned from and similar to the real data. For each replication, phenotypes were simulated under four different models (Fig. 2), capturing the three cases with group effects (monophyletic, polyphyletic and paraphyletic models) and one case with no group effect (nonphyletic model).

To generate the subject phenotypes for each model, the simulated abundance percentages for the OTUs included in the model were summed for each individual and multiplied by a scalar effect size ($\beta = 2$), plus errors randomly drawn from the standard normal distribution. This procedure generated group effects for all modeled OTUs. These values were converted to a binary phenotype by taking subjects above (or below) the median and setting their phenotype value to 1 (or 0). The simulation was repeated 1000 times, each time 500 subjects and 5000 counts per subject were simulated.

We applied SMART-scan, SVA, stepwise regression, LASSO and CART to the data simulated under a the four models, and compared their statistical power of detecting signal OTUs based on the estimated receiver operating characteristic (ROC) curves (Fig. 3), which present the power of the five methods at different levels of type 1 error (from 0 to 0.15). Because a higher ROC curve indicates a better power, it is can be seen from Fig. 3 that SMART-scan has higher power than other four methods when there are group effects (Fig. 3a–c) and its power drops if there is no group effect (Fig. 3d). Among the other four methods, CART produces relatively higher power for all data models. Compared with SVA, the two variable section methods, stepwise regression and LASSO, have no gain in statistical power (although they might outperform SVA in terms of

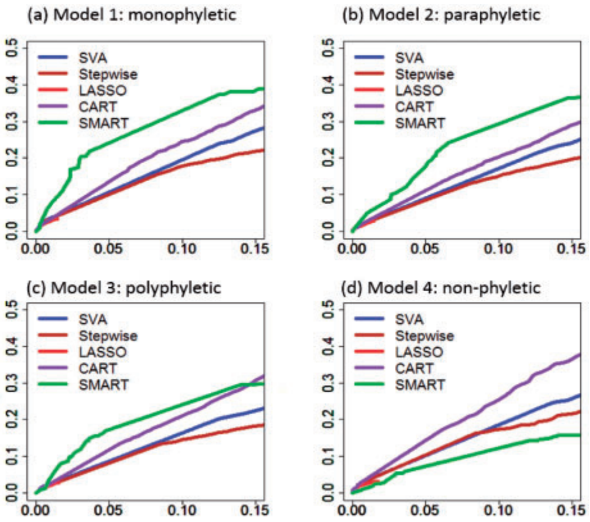

Fig. 3. ROC curves of SVA, stepwise regression, LASSO, CART and SMART-scan. Power (y-axis) and type 1 error rate (x-axis) were estimated by applying the five methods to the data simulated under (**a**) model 1, (**b**) model 2, (**c**) model 3 and (**d**) model 4 for a tree with 287 OTU (bacterial genus) variables (defined in Fig. 2). Each curve was estimated by taking the average power and type 1 error rates over 1000 simulated replications, at a series of different *P* value cutoffs. At a given p-value cutoff, power and type 1 error were estimated, respectively, as the portion of the predictor OTUs (labeled in solid black in Fig. 2) and the portion of the null OTUs (labeled in gray in Fig.2) that were identified to be significant by a method. For each replication and each method, *P* values were calculated by a logistic regression based on the final model (SMART-scan assigns the same *P* values for the OTUs within the same groups in the final group model, SVA assigns p-values for all individual OTUs when stepwise regression, LASSO and CART assign *P* values for only the selected OTUs in the final models). Since power is plotted against observed type 1 error, not p-value cutoff (i.e., expected type 1 error), inflation of *P* value has no effect on the ROC curves

variable selection for the purpose of prediction). Due to the penalty for parameter numbers, LASSO tends to have a good control of false positives and its power is very similar to stepwise regression at small type I error levels.

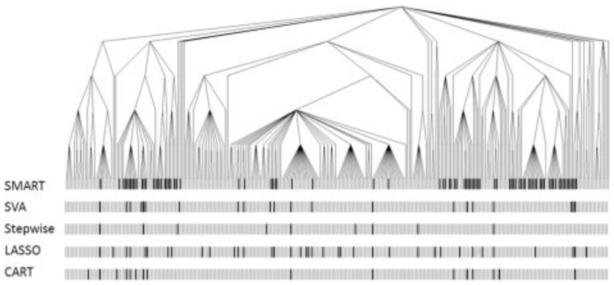

Fig. 4. Diet-associated bacterial genera identified by SMART-scan, SVA.stepwise regression and LASSO. The upper part of the figure shows the taxonomic tree of the 287 genera with informative sequence abundance in the experiment. The bars at the lower area represent individual genera, significant ones (adjusted $P < 0.01$) for SVA and SMART-scan and selected ones by stepwise regression, LASSO and CART are colored in solid black. The five rows of bars indicate five different methods: SMART-scan, SVA, stepwise regression, LASSO and CART

3.2 Application

We used a dataset characterizing the gut microbiome of vervet monkeys (*Chlorocebus pygerythrus*) under two different diets to demonstrate the application of SMART-scan. The production of this data has been described previously (Jasinska *et al.*, 2013) and details of dietary compositions are described elsewhere (Jorgensen *et al.*, 2013; Voruganti *et al.*, 2013). The dataset contains the 16S rRNA gene sequences of gut microbiomes sampled from 179 vervet monkey subjects, of which 68 animals were treated with a fatty diet and 111 with normal diet. Rectal swabs were obtained from the animals and DNA was extracted using the MOBIO PowerSoil kit, and the 16S rRNA genes were amplified (16S rRNA gene variable regions V3–V5) and subjected to pyrosequencing using the 454 Life Sciences sequencing platform. Sequences were compared with the RDP database after quality filtering and the taxa, and their abundances found in each sample were tallied. These procedures were those used in the Human Microbiome Project (2012a, b). These sequences were assigned to a taxonomic tree of 287 bacterial genera, 156 families, 72 orders, 40 classes and 18 phyla (Fig. 2) via the RDP reference alignment, and then the sequence count of each taxon, divided by the total number of sequence reads (averagely 7078 reads per subject), was used as estimates of relative abundance. Animals were designated as being on the normal diet (0) or on the fatty diet (1), analogous to a case-control design. The microbiome dataset we used has been submitted to the NCBI BioProject database with an accession ID PRJNA259274.We applied SMART-scan, SVA, stepwise regression, LASSO and CART to the vervet monkey data, to investigate the effects of a fatty diet on microbiome composition. Out of 287 bacterial genera included in the analysis, SVA identified 21 genera with Bonferroni corrected $P < 0.01$, stepwise regression selected 9 genera, LASSO assigned nonzero effect coefficients to 37 genera, CART chose 16 genera for decision tree building and SMART-scan detected 23 groups of 83 genera with significant *P* values (<0.01) obtained through 1000 permutations for multiple test correction (Fig. 4). A significant feature of SMART-scan is that it identifies two large groups of genera, both with permutation adjusted $P < 0.001$ (here a more accurate *P* value below 0.001 is not reported because only 1000 permutations were used); one group includes 41 genera clustered within the branch of phylum *Proteobacteria* and another group 19 genera from phylum *Bacteroidetes*. In this case, although the results are still needed to be validated, we demonstrate that the application of SMART-scan can reveal important group clustered association features and provide a valuable clue as to the relevant part of the phylogenetic tree associated with microbiome changes in response to diet, which otherwise would not have been statistically detected using other non-tree aware approaches.

4 Discussion

Analysis of microbial communities is rapidly gaining momentum in population studies for its potential as an important biological factor influencing human health. While the hypothesis of microbiome involvement in a variety of phenotypes is appealing, it is not clear what feature of the microbiome may be relevant to a particular host outcome. One way to explore the role of the microbiome is to assess the bacterial species' presence and their relative abundance by targeted sequencing of 16S rRNA with bioinformatics assignment to known classifications. Statistical methods to test the association between these features of microbiome composition with outcome are challenged by the high dimensional nature of the data, coupled with relatively low sample sizes, and the ensuing multiple comparison issues. The situation is further exacerbated because of correlations and interactions between species and genera due to shared ancestry.

All these together make regular variable selection methods suffer from low power.

Microbiome data, however, have a unique feature, i.e. the composition, assessed by sequence analysis, can be organized into a hierarchically structured phylogenetic tree. Because such a tree provides the information on the similarity between taxonomic groups, it can be utilized to significantly improve model selection (Templeton *et al.*, 2005). Based upon this idea, we have developed a tree-based model selection method, SMART-scan, and demonstrated its advantages over other methods. In SMART-scan, we leverage the phylogenetic information in a model selection approach to obtain a more powerful test of association. To realize the benefit of SMART-scan, it is assumed that phylogenetically close taxa have more similar effects than those that are distant. This model appears to be supported by our analysis of the vervet monkey diet experiment, in which of a large group of 41 genera in a phylum was identified to be affected by diet.

Importantly, SMART-scan is not only a useful tool for microbiome analysis, but also provides a strategy and framework for mining high-dimensional structured data, and therefore has applications in other situations. It can be easily extended to other model selection problems in any situation that variables are hierarchically structured and close variables share similar group effects. For example, it can be used to identify genes or gene groups that are associated with phenotypes, using sequence or expression data. In this type of analysis, genes can be treated as variables to be selected and organized into a hierarchical structure based on established pathway or gene set information.

The key feature of SMART-scan is variable grouping with the aid of known variable structure. This feature makes SMART-scan differ from many other multivariate analysis techniques. Unlike the variable selection in stepwise regression, LASSO and CART through including and excluding variables in models, SMART-scan reorganizes variables into groups. Instead of selecting variables, it selects the way of grouping variables. An advantage of variable grouping is that different variables with similar or identical effects are summarized into a single group variable, therefore, the total numbers of variables in the final model could be substantially reduced. From this perspective, SMART-scan can be viewed as a technique of data dimensionality reduction. Compared with other dimensionality reduction methods, such as principal components analysis (PCA) and multidimensional scaling (MDS), SMART-scan has clearer biological meaning and facilitates interpretation. For example, when it is difficult to biologically understand a PCA from a linear combination of hundreds of OTU variables with different loads, we may clearly define a group of 20 OTUs from the same bacterial family in SMART-scan and then investigate the association between them and a clinical outcome. Although SMART-scan utilizes tree-like hierarchical structure, it is very different from hierarchical linear models (HLMs) (also called multilevel models or nested models), because HLMs are for modeling the structure of subjects, not of predictor variables. SMART-scan is also different from classification and regression tree (CART) analysis. CART is a widely-used machine learning method which models the data through a decision tree. A decision tree used by CART is very different from the phylogenetic tree used by SMART-scan in many aspects. It does not necessarily reflect biological phylogeny or taxonomy. Unlike a phylogenetic tree describing the hierarchical structure of variables, a decision tree is a predictive model describing how to use predictors in a treelike ordered Boolean way to predict an outcome. The phylogenetic tree is known and used by SMART-scan as a prior to help grouping and summarizing variables, while a decision tree is unknown and needs to be learned by CART in a data-driven, predictive manner.

These differences, however, do not mean that SMART-scan has no connection with other methods, instead, SMART-scan can be easily incorporated with other methods. For example, after a large number of variables are grouped and summarized by SMART-scan into fewer variables, other methods (e.g. stepwise regression, PCA or CART) can be applied to the new variables to refine the analysis.

Overall, SMART-scan is flexible in terms of data modeling, because the AIC goodness-of-fit makes it applicable to any data type for which likelihood-based models can be developed. It can be applied to quantitative, qualitative and survival phenotypes, with or without covariates, complex interactions and random effects. Currently we have implemented SMART-scan in R, which can handle both quantitative and binary phenotypes, more extensions will be developed and added.

It should be noted that, when a *P* value cutoff is required to determine the significance in an application, directly calculated *P* values in the final model of SMART-scan could be inflated, because the variable grouping and model selection procedure tends to produce results deviating from the null distribution. This is a common issue of all model selection methods, and usually can be adjusted by a permutation test, which can be implemented by users through a standard, random permutation of phenotype data (Supplementary Fig. S1).

Another caveat to the application of SMART-scan is that the result of an analysis and its interpretation rely on an estimated phylogenetic tree, and this estimated tree can be subject to ambiguity due to homoplasy because of parallel or convergent mutations and/or horizontal gene transfer. Concerning tree ambiguity, earlier work with nested clade analysis and TreeScan, techniques that also use a phylogenetic tree to generate a statistical design and that motivated SMART-scan, showed that the statistical design that emerges from using an estimated tree is actually robust to much potential ambiguity within that estimated tree and does not require a fully resolved tree (Templeton *et al.*, 2005; Templeton and Sing, 1993). For example, most homoplasy occurs on long branches that separate phyla. However, SMART-scan does not require that the phylogenetic relationships among phyla be fully resolved; rather, SMART-scan only requires that the basic topology be an accurate one and does not require accurate branch lengths, exact interconnections or rooting of the tree. We used the taxonomic classification as the basis of our tree in this paper, and a recent estimator of the bacteria phylogeny based on whole genomes revealed 'a general agreement between the taxonomic classification and phylogenetic distribution of these genomes' (Chai *et al.*, 2014). Within phyla, homoplasy is less of a problem and the estimated trees are more accurate, so one great advantage of just using the tree to generate a statistical design is that it makes homoplasy much less of a problem in inference of trait associations as opposed to inference of phylogeny. However, one aspect of horizontal gene transfer could be important in our inference structure. In Section 3.2, we used sequence data to estimate the bacterial abundance (down to the genus level) by counting the sequence reads that were mapped to a set of reference sequences with known, predefined taxonomy. When there are horizontal gene transfer events causing homoplasy between genera, the abundance estimation could be biased, because a transferred gene sequence in genus A could be aligned back to its original location (genus B) in the tree, causing it to be counted for the abundance of genus B and not genus A. Accordingly, we may interpret the abundance in a slightly different way: it is not strictly generic abundance, but rather sequence-oriented abundance that complies with the reference sequences. Therefore, the actual goal of Section 3.2 can be understood as the identification of association between diet and sequences clustered based on their similarity, not association between diet and strictly biological abundances. Alternatively, an unsupervised clustering tree based on only sequence similarity (not taxonomy reference alignment) can also be used for the SMART-scan analysis. Of course, when biological abundances are of interest, other analytical strategies may need to be adopted, such as using data at higher taxonomic levels or using subtrees with less homoplasy.

Acknowledgements

We thank Kathie Mihindukulasuriya, Yanjiao Zhou and Erica Sodergren for technical assistance and access to the vervet monkey data.

Funding

The National Institute of Health (NIH) (1R01DK8925601 to I.B.B) and (5R01DK075681 to I.B.B).

Conflict of Interest: none declared.

References

Ahn,J. *et al.* (2013) Human gut microbiome and risk for colorectal cancer. *J. Natl. Cancer Inst.*, **105**, 1907–1911.

Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol.*, **26**, 32–46.

Caporaso,J.G. *et al.* (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*. **26**, 266–67.

Chai,J. *et al.* (2014) Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. *BMC Evol. Biol.*, **14**, 207.

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, **28**, 2106–2113.

DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.

Dey,N. *et al.* (2013) Association of gut microbiota with post-operative clinical course in Crohn's disease. *BMC Gastroenterol.*, **13**, 131.

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.

Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences*. University of Washington, Seattle.

Felsenstein,J. (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am. Nat.*, **171**, 713–725.

Garcia,T.P. *et al.* (2013) Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics*, **30**, 831–837.

Hill,T.C. *et al.* (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol. Ecol.*, **43**, 1–11.

Holmes, I. *et al.* (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, **7**, e30126.

Human Microbiome Project, C. (2012a) A framework for human microbiome research. *Nature*, **486**, 215–221.

Human Microbiome Project, C. (2012b) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Jasinska, A.J. *et al.* (2013) Systems biology of the vervet monkey. *ILAR J.* **54**, 122–143.

Jorgensen, M.J. *et al.* (2013) Effects of a Western-type diet on plasma lipids and other cardiometabolic risk factors in African green monkeys (Chlorocebus aethiops sabaeus). *J. Am. Assoc. Lab. Anim. Sci.*, **52**, 448–453.

La Rosa,P.S. *et al.* (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, **7**, e52078.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Ludwig,W. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.

Mantel,N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.*, **27**, 209–220.

Pruesse, E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Segata,N. *et al.* (2012) Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.*, **13**, R42.

Sun,Y. *et al.* (2010) Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Res.*, **38**, e205.

Templeton,A.R. and Sing,C.F. (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Templeton,A.R. *et al.* (2005) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, **169**, 441–453.

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

Voruganti,V.S. *et al.* (2013) Significant genotype by diet (G x D) interaction effects on cardiometabolic responses to a pedigree-wide, dietary challenge in vervet monkeys (Chlorocebus aethiops sabaeus). *Am. J. Primatol.*, **75**, 491–499.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.

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

Wooley,J.C. *et al.* (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.