Alshawaqfeh *et al. BMC Bioinformatics* (2017) 18:328
DOI 10.1186/s12859-017-1738-1

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Reliable Biomarker discovery from Metagenomic data via RegLRSD algorithm

Mustafa Alshawaqfeh[1], Ahmad Bashaireh[1], Erchin Serpedin[1*] and Jan Suchodolski[2]

Abstract

Background: Biomarker detection presents itself as a major means of translating biological data into clinical applications. Due to the recent advances in high throughput sequencing technologies, an increased number of metagenomics studies have suggested the dysbiosis in microbial communities as potential biomarker for certain diseases. The reproducibility of the results drawn from metagenomic data is crucial for clinical applications and to prevent incorrect biological conclusions. The variability in the sample size and the subjects participating in the experiments induce diversity, which may drastically change the outcome of biomarker detection algorithms. Therefore, a robust biomarker detection algorithm that ensures the consistency of the results irrespective of the natural diversity present in the samples is needed.

Results: Toward this end, this paper proposes a novel Regularized Low Rank-Sparse Decomposition (RegLRSD) algorithm. RegLRSD models the bacterial abundance data as a superposition between a sparse matrix and a low-rank matrix, which account for the differentially and non-differentially abundant microbes, respectively. Hence, the biomarker detection problem is cast as a matrix decomposition problem. In order to yield more consistent and solid biological conclusions, RegLRSD incorporates the prior knowledge that the irrelevant microbes do not exhibit significant variation between samples belonging to different phenotypes. Moreover, an efficient algorithm to extract the sparse matrix is proposed. Comprehensive comparisons of RegLRSD with the state-of-the-art algorithms on three realistic datasets are presented. The obtained results demonstrate that RegLRSD consistently outperforms the other algorithms in terms of reproducibility performance and provides a marker list with high classification accuracy.

Conclusions: The proposed RegLRSD algorithm for biomarker detection provides high reproducibility and classification accuracy performance regardless of the dataset complexity and the number of selected biomarkers. This renders RegLRSD as a reliable and powerful tool for identifying potential metagenomic biomarkers.

Keywords: Biomarker detection, Metagenomics, Matrix decomposition, Alternating direction method of multipliers, Augmented Lagrangian

Background

Thanks to the progress witnessed by the high-throughput sequencing technologies, large-scale investigation of bacterial collectivities has become possible by means of metagenomic approaches. This large-scale analysis lead to the discovery of bacterial groups that could not be analyzed through the conventional cultivation-based methods (90% of microbes are not recognized yet and not cultivable [1, 2]). In addition to bacterial composition, metagenomic techniques employed the whole-metagenome shotgun sequencing methods to infer the functional role of microbial colonies [3, 4].

Recently, several metagenomic studies have pointed out that the distortion of the normbiosis state of bacterial communities is a key player in the progression of many diseases such as obesity [5–7], diabetes [8], inflammatory bowel disease (IBD) [9], and cancer [10, 11]. These findings suggest employing microbes as possible biomarkers for the health status and certain diseases of the host. Currently, the determination of microbial biomarkers is carried out by finding the operational taxonomic units

*Correspondence: eserpedin@tamu.edu
[1]Bioinformatics and Genomic Signal Processing Lab, ECEN Dept., Texas A&M University, 77843-3128, College Station, TX, USA
Full list of author information is available at the end of the article

(OTUs), whose corresponding abundances differentiate for samples pertaining to distinct phenotypes.

Biomarker detection is crucial to understand disease development and design antibiotic and/or probiotic therapies. Mathematically, the task of biomarker identification can be formulated as determining the most revealing features that can differentiate multiple sets of samples or conditions (i.e., various stages of a disease, different categories of diseases, etc.). The methods proposed in literature to address the biomarker discovery problem can be classified into two categories: machine learning (pattern recognition) methods and statistical methods, respectively.

In general, the statistical approaches tackle the problem by using a statistical hypothesis test to calculate the statistical significance (i.e., p-value) of each feature. Then, the features associated with p-values lower than a well-selected level are declared as potential biomarkers. A major issue linked with the statistical-based methods is the multiple comparisons problem, which is commonly solved by substituting the p-values with the corresponding false discovery rates (FDRs). Metastats [12] and LEfSe [13] are the current standard approaches that belong to this category. Specifically, Metastats utilizes the permutation t-test and the exact Fisher's test for non-sparse and sparse features, respectively [12]. On the other hand, to improve the robustness of biomarker discovery, LEFSe relates the statistical study with the impact of size estimation [13]. In particular, LEFSe exploits the Kruskal-Wallis and Wilcoxon-Mann-Whitney detection algorithms for class and subclass comparative studies, respectively.

In the machine learning framework, the problem of detecting the biomarkers is formulated as a feature determination task. The filtering methods are the most widely adopted approaches for biomarker detection. In filtering methods, each OTU is assigned a score based on the relevance between its abundance levels across the samples and the class labels of the samples. The operational taxonomic units that present the largest scores are declared as potential biomarkers. This scoring process is carried out one by one for each OTU and separately of the other OTUs. Therefore, filtering methods are computationally fast and easily interpretable. However, the individual ranking ignores the inter-dependencies among different variables.

Contrary to the individual ranking, the feature transformation-based methods try to generate more revealing features where each newly detected feature is depedent of all the original features. Considering all the initial characteristics in the construction of new traits accounts for the interactions between OTUs. Transformation approaches are divided broadly into two categories based on whether the labels of the samples are considered in the transformation process. These categories are the supervised and unsupervised approaches. Linear discriminant analysis (LDA) and partial least-squares (PLS) represent the two most employed supervised approaches. On the other hand, the principal component analysis (PCA) presents itself as one of the most remarkable unsupervised methods.

Identifying the most discriminating features in metagenomic datasets is a challenging task. One major challenge is that the number of biomarkers might be much larger than the number of available samples, a condition that it is commonly termed as the 'high dimension low-sample size (HDLSS)' problem. The HDLSS problem is also associated with serious analytical challenges [14, 15]. In addition, metagenomic analysis presents its own challenges such as: (i) metagenomic-specific artifacts such as sequencing errors and chimeric reads [16, 17], (ii) high dynamics of the bacterial populations due to the complex interactions with the host [18] and between its members [19–21], and (iii) inter-subject variability. For example, the results of [6] show that the gut microbiota of twins differ significantly.

These challenges point to a severe inconsistency issue that blocks the current biomarker identification methods from selecting the true biomarkers. For example, the authors of [22] reported that out of the 70 genes that were suggested as potential biomarkers for breast cancer by the two gene expression studies [23, 24], only three genes were found to be common. Therefore, developing a robust biomarker detection algorithm that ensures the reproducibility of the outcomes obtained from biological data plays a critical role in infering correct biological statements and making use of these results in good clinical decisions.

Toward this end, we propose herein paper the Regularized Low Rank-Sparse Decomposition (RegLRSD) algorithm for biomarker detection. RegLRSD formulates the biomarker discovery problem as a matrix decomposition problem and provides an efficient solution for this decomposition. In particular, RegLRSD models the bacterial abundance data as the superposition of a sparse matrix and a low-rank matrix. The motivation for this is due to the fact that most of microbes do not play any role. Hence, the abundance profiles of these uninformative bacteria do not vary between samples associated with different phenotypes. Therefore, considering their abundance profile as a low-rank matrix is natural. In addition, few microbes may be relevant to the biological condition under study. Consequently, the abundance profiles of these relevant microbes are expected to vary significantly between the different phenotypes. Therefore, modeling these informative bacteria as a sparse matrix is legitimate.

To improve the accuracy of extracting the low-rank and sparse matrices, we exploit the prior knowledge that the abundance profiles of non-informative bacteria do not exhibit significant variation. This is achieved by adding a smoothness constraint on the recovered low

rank matrix. The RegLRSD algorithm presents several advantages. First, RegLRSD improves the reproducibility performance because of the following traits: (i) RegLRSD incorporates prior knowledge in the detection process, which constrains the analysis. Consequently, this mitigates the conventional challenges associated with the HDLSS nature of metagenomic data. (ii) The multivariate nature of RegLRSD algorithm accounts for the complex interactions between the members of the bacterial community. This contrasts the univariate-based methods (i.e., statistical hypothesis testing and filtering techniques) that ignore such sophisticated relationships between bacteria. Second, the proposed matrix decomposition formulation is convex. This provides several benefits such as: (i) global optimality, (ii) efficient solvers, and (iii) flexibility to add convex constraints without affecting the convex structure of the problem. Third, unlike feature transformation-based algorithms, the output of RegLRSD is easily interpretable in the sense that it keeps the features in their original domain.

This paper also sheds light into the design of an evaluation protocol which provides a fair and an accurate assessment of the efficiency of a biomarker detection algorithm. The absence of the "ground truth" (i.e., no absolute knowledge of the true biomarkers) prevents the objective evaluation of the biomarker detection methods. Therefore, the assessment criteria and comparisons have to be conducted with great care to make sure that all the existing prior knowledge about the true markers is taken into account.

Methods

Low rank-sparse model of metagenomic data

Consider the matrix $\mathbf{D} \in \Re^{p\times n}$ of bacterial abundance data, each line of $\mathbf{D}$ denotes the relative abundance of an OTU in all the n samples, and each column stands for the abundance values of all the p OTUs in one sample. In general, $p \gg n$. Therefore, it represent a challenging high-dimensional small-sample size problem. The backbone of our approach is to capture the differentially and non-differentially abundant OTUs via a sparse matrix and low-rank matrix, respectively. In particular, most of the bacterial groups do not play any role in the considered biological system. Thus, these inappropriate OTUs are expected to exhibit high abundance levels that do not change significantly between two different phenotypes. Therefore, it makes perfect sense to model their abundance-level matrix as a low-rank matrix (represented by matrix $\mathbf{L}$). Also, the abundance levels of the few key OTUs might present relevant changes between the two phenotypes. Such a condition will be captured by means of a sparse matrix (in our case, the matrix $\mathbf{S}$). Mathematically,

$$\mathbf{D} = \mathbf{L} + \mathbf{S}. \tag{1}$$

Extracting the sparse matrix via RegLRSD

Exploiting the low rank-sparse decomposition model of the bacterial abundance profiles (1), identifying potential biomarkers boils down to a matrix decomposition problem, with the aim of extracting the sparse matrix. This decomposition can be cast mathematically as the following optimization:

$$\begin{aligned} &\text{minimize } \ \text{rank}(\mathbf{L}) + \lambda\|\mathbf{S}\|_0 \\ &\text{subject to } \ \mathbf{D} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{2}$$

where $\|\mathbf{S}\|_0$ denotes the l_0-norm of the matrix $\mathbf{S}$, which by definition is equal to the number of nonzero elements in $\mathbf{S}$. Problem (2) is commonly known as the robust PCA (RPCA) problem. This formulation of RPCA, given by (2), is highly non-convex because of the combinatorial optimization required by the rank operator and the l_0-norm. However, the authors in [25, 26] pointed out that under general conditions, one *exactly* estimate both components (i.e., low rank and sparse matrices) by carrying out a convex optimization, referred to as the Principal Component Pursuit (PCP). This convex formulation is based on recent theories and results that show: (i) the l_1 norm represents the closest convex approximation of the l_0-norm, and minimizing l_1-norm yields the sparsest solution to underdetermined linear systems [27], (ii) the nuclear norm provides a tight approximation of the matrix rank operator and minimizing the nuclear norm provides the lowest rank solution under wide assumptions [28]. Mathematically, PCP is expressed as

$$\begin{aligned} &\text{minimize } \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \\ &\text{subject to } \ \mathbf{D} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{3}$$

where λ represents a positive regularization factor that monitors the degree of sparseness and smoothness in $\mathbf{S}$ and $\mathbf{L}$, respectively. Variable $\|\mathbf{L}\|_*$ stands for the nuclear norm of $\mathbf{L}$ and is equal to the sum of the singular values. Finally, the notation $\|\mathbf{S}\|_1$ denotes the l_1 norm of $\mathbf{S}$, and it is defined as the summation of the absolute values of the matrix elements.

In an attempt to enhance the estimation accuracy of $\mathbf{S}$ and $\mathbf{L}$, we extend the formulation in (3) by adding a penalty term in order to enforce the smoothness of each row of $\mathbf{L}$. This penalty term incorporates the prior knowledge that the abundance profiles of non differentially abundant OTUs are smooth. In this paper, the first order difference (FOD) is adopted as a measure of smoothness, which is defined as:

$$\|\mathbf{X}\|_{FOD} = \sum_j \|\mathbf{F}\mathbf{x}_j\|_1, \tag{4}$$

where $\mathbf{x}_j$ denotes the j^{th} column of $\mathbf{X}$, and $\mathbf{F}$ represents the first order difference operator defined as:

$$\mathbf{F} = \begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{5}$$

Thus, the RegLRSD algorithm aims to untie the optimization problem:

$$(\mathbf{L}^*, \mathbf{S}^*) = \arg\min_{\mathbf{L},\mathbf{S}} \left\{ f(\mathbf{D}, \mathbf{L}, \mathbf{S}) = \frac{1}{2}\|\mathbf{D} - \mathbf{L} - \mathbf{S}\|_F^2 + \alpha\|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{l}_i^T\|_1 \right\}, \tag{6}$$

where $\mathbf{l}_i^T$ stands for the i^{th} row of $\mathbf{L}$. One key advantage of this formulation is that that the optimization problem (6) is convex. The above-mentioned convex optimization formulation yields several benefits: (i) it enables a global optimal solution, (ii) it enables utilizing the well-established theory and tools for solving convex optimization problems, and (iii) it allows the luxury to take into account extra convex constraints to capture better the existing prior information. However, direct application of generic convex solvers may not be feasible due to the high dimensional nature of our problem. For example, interior point methods exhibit high order complexity. Moreover, there is no approach available to determine the jointly optimal solution for the optimization (6). Therefore, herein paper we consider an efficient alternating-based algorithm to carry out (6). The alternating-minimization approach first optimizes $f(\mathbf{L}, \mathbf{S})$ with respect to $\mathbf{S}$ (matrix $\mathbf{L}$ is considered constant), and then it optimizes $f(\mathbf{L}, \mathbf{S})$ with respect to $\mathbf{L}$ (matrix $\mathbf{S}$ being considered a fixed constant). In particular, it adopts the following updating steps:

$$\mathbf{S}^{(k)} = \arg\min_{\mathbf{S}} f(\mathbf{L}^{(k-1)}, \mathbf{S}) \tag{7}$$

$$\mathbf{L}^{(k)} = \arg\min_{\mathbf{L}} f(\mathbf{L}, \mathbf{S}^{(k)}). \tag{8}$$

This strategy utilizes the fact that the two sub-problems (7) and (8) admit efficient solutions. In particular, the problem in (7) can be reformulated as follows:

$$\mathbf{S}^{(k)} = \arg\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{D} - \mathbf{L}^{(k-1)} - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1. \tag{9}$$

Problem (9) admits the following closed form solution:

$$\mathbf{S}^{(k)} = \mathcal{S}_\lambda(\mathbf{D} - \mathbf{L}^{(k-1)}), \tag{10}$$

where $\mathcal{S}_\tau : \Re \rightarrow \Re$ denotes the *shrinkage operator*, expressed as:

$$\mathcal{S}_\tau(x) = \text{sgn}(x)\max(|x| - \tau, 0), \tag{11}$$

and where $\tau \geq 0$ denotes the threshold level. In the case of a matrix, the shrinkage operator will be applied onto each constituent element of the matrix. The problem in (8) can be cast as:

$$\mathbf{L}^{(k)} = \arg\min_{\mathbf{L}} \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{l}_i^T\|_1. \tag{12}$$

The current formulation of the optimization problem in (12) is neither in a format that admits a closed-form expression as (7) nor in the format of a well-established problem that admits an efficient solution. Moreover, relying on generic convex techniques to solve (12) may not be efficient. The difficulty exhibited by this minimization problem arises from the combination of the two non-smooth terms $\|\mathbf{L}\|_*$ and $\sum_{i=1}^{p} \|\mathbf{F}\mathbf{l}_i^T\|_1$. Therefore, we propose to reformulate (12) by introducing an additional variable and constraint to separate these two terms. Adding this auxiliary variable enables the decomposition of (12) into two subproblems that can be solved efficiently. The first subproblem is the *nuclear-norm regularized least-squares* (LS) optimization problem which presents a closed-form solution [29]. The second problem can be recast as the *total variation denoising* problem [30], which presents an efficient solution [31]. In particular, (12) is reformulated as:

$$\begin{aligned} (\mathbf{L}, \mathbf{Y}) = \arg\min_{\mathbf{L},\mathbf{Y}} \ & \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} + \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* \\ & + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{y}_i^T\|_1, \\ \text{subject to } & \mathbf{Y} = \mathbf{L}, \end{aligned} \tag{13}$$

where $\mathbf{y}_i^T$ stands for the i^{th} row of the auxiliary variable $\mathbf{Y}$. To solve (13), we make use of the alternating direction method of multipliers (ADMM) [31]. In general, the ADMM algorithm converts the constrained optimization problem into an unconstrained optimization problem with a novel objective that it is referred to as the augmented Lagrangian. The augmented Lagrangian associated with the optimization (13) is:

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}, \mathbf{Z}) = & \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} + \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* \\ & + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{y}_i^T\|_1 + \langle \mathbf{Z}, \mathbf{L} - \mathbf{Y} \rangle \\ & + \frac{\rho}{2}\|\mathbf{L} - \mathbf{Y}\|_F^2, \end{aligned} \tag{14}$$

where $\mathbf{Z}$ represents the Lagrange multiplier matrix. Thus, the ADMM formulation of (13) is given by:

$$(\mathbf{L}, \mathbf{Y}, \mathbf{Z}) = \arg\min_{\mathbf{L},\mathbf{Y},\mathbf{Z}} \mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}, \mathbf{Z}). \tag{15}$$

The ADMM solution of (15) is of recursive nature. Each recursion, in particular the r-th iteration, assumes the updates:

$$\begin{aligned}\mathbf{L}^{(r)} &= \arg\min_{\mathbf{L}} \frac{1}{2}\|\mathbf{D}-\mathbf{S}^{(k)}-\mathbf{L}\|_F^2 \\ &\quad + \alpha\|\mathbf{L}\|_* + \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L}-\mathbf{Y}^{(r-1)} \right\rangle \\ &\quad + \frac{\rho}{2}\|\mathbf{L}-\mathbf{Y}^{(r-1)}\|_F^2,\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbf{Y}^{(r)} &= \arg\min_{\mathbf{Y}} \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L}^{(r)}-\mathbf{Y} \right\rangle \\ &\quad + \frac{\rho}{2}\|\mathbf{L}^{(r)}-\mathbf{Y}\|_F^2 + \beta\sum_{i=1}^{p}\|\mathbf{F}\mathbf{y}_i^T\|_1,\end{aligned} \tag{17}$$

$$\mathbf{Z}^{(r)} = \mathbf{Z}^{(r-1)} + \rho(\mathbf{L}^{(r)}-\mathbf{Y}^{(r)}) \tag{18}$$

Remark 1 *For any arbitrary vectors* $\mathbf{u},\mathbf{v} \in \Re^n$*, and scalars* $a,b \in \Re$*, the following relation holds:*

$$\langle a\mathbf{v}+b\mathbf{u},\mathbf{u}\rangle = b\left\|-\frac{a}{2b}\mathbf{v}-\mathbf{u}\right\|_F^2 - \frac{a^2}{4b}\|\mathbf{v}\|_F^2. \tag{19}$$

Based on Remark-1, the problem in (16) is recast as:

$$\begin{aligned}\mathbf{L}^{(r)} &= \arg\min_{\mathbf{L}} \alpha\|\mathbf{L}\|_* \\ &\quad + \frac{1+\rho}{2}\left\|\frac{\mathbf{D}-\mathbf{S}^{(k)}+\rho\mathbf{Y}^{(r-1)}-\mathbf{Z}^{(r-1)}}{1+\rho}-\mathbf{L}\right\|_F^2\end{aligned} \tag{20}$$

According to [29], problem (20) admits the following closed form solution:

$$\mathbf{L}^{(r)} = \mathcal{D}_{\frac{\alpha}{1+\rho}}\left(\frac{\mathbf{D}-\mathbf{S}^{(k)}+\rho\mathbf{Y}^{(r-1)}-\mathbf{Z}^{(r-1)}}{1+\rho}\right), \tag{21}$$

where $\mathcal{D}_\tau$ is the *singular value shrinkage* operator defined by:

$$\mathcal{D}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{D}_\tau(\Sigma)\mathbf{V}^T, \quad \mathcal{D}_\tau(\Sigma) = \mathrm{diag}(\{\sigma_i-\tau\}_+) \tag{22}$$

where $\mathbf{U}$, $\mathbf{V}$, and σ_i stand for the left singular vectors, right singular vectors and singular values of $\mathbf{X}$, respectively, and the notation $(x)_+$ denotes the positive part of x (i.e., $(x)_+ = \max(0,x)$). In other words, $D_\tau(\mathbf{X})$ employs a soft-thresholding operation onto the singular values of $\mathbf{X}$, shifting these towards zero. This is the reason why this transformation it is also referred to as the *singular value shrinkage* operator.

Considering Remark-1, problem (17) is recast as:

$$\begin{aligned}\mathbf{Y}^{(r)} &= \arg\min_{\mathbf{Y}} \frac{\rho}{2}\left\|\frac{\mathbf{Z}^{(r-1)}+\rho\mathbf{L}^{(r)}}{\rho}-\mathbf{Y}\right\|_F^2 \\ &\quad + \beta\sum_{i=1}^{p}\|\mathbf{F}\mathbf{y}_i^T\|_1.\end{aligned} \tag{23}$$

The rows of $\mathbf{Y}$ are updated separately according to the optimization:

$$\begin{aligned}\mathbf{y}_i^{T(r)} &= \arg\min_{\mathbf{y}} \frac{\rho}{2}\left\|\frac{\mathbf{z}_i^{T(r-1)}+\rho\mathbf{l}_i^{T(r)}}{\rho}-\mathbf{y}\right\|_F^2 \\ &\quad + \beta\|\mathbf{F}\mathbf{y}\|_1,\end{aligned} \tag{24}$$

where $\mathbf{z}_i$ and $\mathbf{l}_i$ are the i^{th} rows of $\mathbf{Z}$ and $\mathbf{L}$, respectively. Problem (24) is often called the total variation denoising problem [30], and it admits an efficient solution via ADMM as described in Section 6.4.1 in [31]. Alternatively, problem (24) can be cast as a special case of the Fused Lasso Signal Approximator (FLSA), which can be properly addressed via the subgradient finding algorithm (SFA) [32].

The RegLRSD algorithm is summed up via Algorithm 1.

Algorithm 1: RegLRSD algorithm to solve the regularized low rank-sparse matrix decomposition problem (6).

```
Input  : D
while not convereged do
    update S^k using Eq. 10;
    while not convereged do
        update L^r using Eq. 21;
        update Y^r by solving (24) using ADMM solver
        or FLSA solver;
        update Z^r using Eq. 18;
    end
    L^k ← L^r;
end
Output: L, S
```

Extracting the differentially abundant bacteria via RegLRSD

The proposed approach for biomarkers detection assumes two stages. First, employ RegLRSD to resolve the original bacterial abundance data matrix into a low-rank matrix that models the non-differential abundant bacteria and a sparse matrix that models the differential abundant bacteria. Second, construct a scoring vector as a function of the extracted sparse matrix to rank each OTU (i.e., feature). Then, the m highest scores OTUs are declared as potential bacterial biomarkers.

The reasoning for employing the sparse matrix for extracting the potential biomarkers is that the abundance levels of informative OTUs can be considered to be a sparse perturbation matrix superposed over the low-rank matrix that models the abundance levels of the non-informative microbes (i.e., $\mathbf{D}=\mathbf{L}+\mathbf{S}$). The stronger the variation in the abundance levels of OTUs, the larger the magnitude of the corresponding elements in the sparse

matrix **S**. It is pertinent to mention that the strength of the variation of each OTU between the two phenotypes is determined by the absolute values of the non-zero entries in **S** rather than their exact values. This is because the elements of **S** could be either positive or negative based on the role (i.e., activation or deactivation) played by the microbes. Therefore, the score of the i^{th} OTU is achieved by adding up the absolute values of the elements located on the i^{th} line of **S**. Thus, the scoring vector $\mathbf{s}v$ is expressed as :

$$\mathbf{v} = \left[\sum_{j=1}^{n} |s_{1j}|, \ldots, \sum_{j=1}^{n} |s_{pj}| \right]^T. \quad (25)$$

Parameter selection

RegLRSD algorithm is equipped with four regularization parameters, α, β, λ and ρ that control the impact of the rank (i.e., $\|\mathbf{L}\|_*$), smoothness $\left(i.e., \sum_{i=1}^{p} \|\mathbf{F}\mathbf{l}_i^T\|_1\right)$, sparseness ($i.e.$, $\|\mathbf{S}\|_0$), and fitness $\left(i.e., \|\mathbf{L}-\mathbf{Y}\|_F^2\right)$ penalties in (6) and (14). In order to select the appropriate values for these parameters, we relied on similar models and utilized the recommended settings proposed in literature. For example, the PCP problem (3), which is a pruned variant of the objective of RegLSRD algorithm, was addressed in [26]. In particular, PCP assumes the following objective $\|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_0$. The authors in [26] proved that under mild assumptions, the two matrices **L** and **S** can be recovered with high probability when $\lambda/\alpha = 1/\sqrt{\max\{n,p\}}$. Therefore, in our experiments, we set $\alpha = 1$ and $\lambda = 1/\sqrt{\max\{n,p\}}$.

In what concerns the fitness penalty parameter ρ, which is the single parameter that is associated with the ADMM method, the ADMM technique is known for its robustness to poor selection of its parameter. Specifically, the convergence of ADMM is guaranteed, under broad assumptions, for all positive values of its parameter [33]. Here, we set $\rho = 1$. In addition, herein paper, we set $\beta = 0.1\alpha$.

Implementation and disponibility of the method

The RegLRSD algorithm is carried out in MATLAB and exploits the original codes of the SFA algorithm (i.e., "flsa" function included in the SLEP package [34]) in order to solve the subproblem (24). Therefore, RegLRSD cannot be used for commercial applications without consent from the authors of SFA algorithm and RegLRSD. To support ongoing metagenomic analysis and to extend the utility of RegLRSD for non-MATLAB users, RegLRSD is implemented as a standalone executable software package and is made available at https://sites.google.com/a/tamu.edu/mustafa/software/reglrsd. This package is provided with a graphical interface to enable the user to set the algorithm parameters and to report the detected markers.

Nearest centroid classifier (NCC)

A nearest centroid classifier represents a special case of a distance-based supervised learning approach. The NCC-based classification approach assumes two steps. The first step trains the classifier by exploiting the labeled data (i.e., $\mathbf{d}_i$) to determine the mean (i.e., centroid) of each class. The average vakue of the $k'th$ class (μ_{C_k}) is obtained as follows:

$$\mu_{C_k} = \frac{1}{|N_{C_k}|} \sum_{\mathbf{d}_i \in C_k} \mathbf{d}_i. \quad (26)$$

The second step assigns a test sample ($\mathbf{z}$) to the class that presents a closer centroid. This reduces to the optimization:

$$\hat{C}(\mathbf{z}) = \arg\min_{C_k} dis(\mu_{C_k}, \mathbf{z}), \quad (27)$$

where $dis(\mu_{C_k}, \mathbf{z})$ stands for the distance between the test sample $\mathbf{z}$ and the centroid of the samples associated with the $k'th$ class (μ_{C_k}).

Data description

The abundance levels of the OTUs were generated from filtered 16S rRNA gene sequencing by exploiting the naive Bayesian classifier already implemented in the Ribosomal Database Project (RDP) [35]. The reads that present confidence below 0.8 were rebinned not certain. The per-sample normalized bacterial abundance profiles were collected into a matrix, referred to as the taxonomic relative abundance matrix. RegLRSD algorithm takes this matrix as input. Due to the unsupervised nature of RegLRSD, the sample labels are not necessary.

Dogs with idiopathic inflammatory bowel disease (IBD) dataset

This dataset compares the fecal microbiota between 10 healthy dogs and 12 dogs diagnosed with IBD. The extracted DNA from fecal samples was sequenced by 454-pyrosequencing. OTUs were attributed by making sure at least 97% sequence similarity against the Greengenes reference database [36] using Quantitative Insights Into Microbial Ecology (QIIME) [37]. The sequencing data were stored into the National Center for Biotechnology Information (NCBI)-Sequence Read Archive (SRA) with the registration number SRP040310.

Dogs with exocrine pancreatic insufficiency (EPI) dataset

Three day pooled fecal samples were gathered from 18 healthy dogs and 7 dogs with EPI. Extracted DNA was sequenced by Illumina sequencer, and the generated sequences were analyzed using QIIME to obtain the final OTU table with at least 97% sequence similarity against the Greengenes reference database. The sequences can be

accessed in the NCBI-SRA database under the accession number SRP091334.

Mouse model of ulcerative colitis (UC) dataset

This data set stands for the fecal microbiota of the mice model with UC and control mice. The description of the samples collection, processing and DNA extraction is described in [38]. The microbiota of 20 T-bet$^{-/-}$ x Rag2$^{-/-}$ (UC) and 10 Rag2$^{-/-}$ (control) mice was assessed using 16S data from fecal samples. The taxonomic relative abundance table is publicly available in the Supplementary Material of [13].

Results and discussions

This section presents the comparison of RegLRSD algorithm with the latest existing algorithms over the three metagenomic investigations described in the Material and Methods Section. In particular, the RegLRSD algorithm is contrasted with LEFSe [13] and MetaStats [12] from the statistical biomarker detection algorithms family, MetaBoot [39] and the entropy-based filtering method from the machine learning family. Additionally, RegLRSD is compared with the RPCA algorithm for metagenomic biomarker detection [40] in order to examine the impact of adding the smoothness constraint into the original PCP problem (2).

Evaluation criteria

The competing algorithms were evaluated based on their classification and reproducibility performance. The essence of this evaluation relies on generating a high number of variations in the original dataset. Then, the evaluation metrics are computed by averaging the results obtained over all these different variations as shown Algorithm 2. The details of the evaluation protocol is discussed in the following two subsections.

Algorithm 2: Evaluation protocol for assessing the the reproducibility and classification performance.

Input : $\mathbf{D}$
for $k = 1 : K$ **do**
  divide $\mathbf{D} \in \Re_+^{p\times n}$ into two subsets:
  - Training set: $\mathbf{D}_k^{train} \in \Re_+^{p\times \lceil r.n \rceil}$.
  - Testing set: $\mathbf{D}_k^{test} \in \Re_+^{p\times (n-\lceil r.n \rceil)}$

  apply the biomarker detection algorithm over $\mathbf{D}_k^{train}$
  train the classifier with $\mathbf{D}_k^{train}$
  test the classifier against $\mathbf{D}_k^{test}$
end
compute the average consistency using Eq. 28
compute the average sensitivity, specificity, and accuracy over the K iterations

Reproducibility performance

The reproducibility performance of a biomarker detection algorithm is empirically measured by generating different variations of the original dataset, and comparing the output of the algorithm based on these different variations. The reasoning behind this procedure is that a stable biomarker detection approach must provide alike outcomes in the presence of small variations in the data samples. This requirement is in line with the hopes of biologists that expect that changing the sample size by taking out or including a few samples must not alter dramatically the biomarkers detected by the algorithm.

The evaluation methodology for estimating the reproducibility performance can be formalized as follows. First, divide the original dataset $\mathbf{D} \in \Re_+^{p\times n}$ into two subsets: $\mathbf{D}_k^{train} \in \Re_+^{p\times \lceil r\cdot n \rceil}$ and $\mathbf{D}_k^{test} \in \Re_+^{p\times (n-\lceil r\cdot n \rceil)}$, where $r \in (0, 1)$. This random division is repeated K times, and the sub-index k represents the iteration number. Second, the biomarker detection algorithm is applied on each of the K training subsets. This results in K sets of potential biomarkers (i.e., $\{\mathcal{F}_k\}_{k=1}^K$, where $\mathcal{F}_k$ denotes the set of identified markers when applying the algorithm over $\mathbf{D}_k^{train}$). Third, the pairwise similarity between the $K(K-1)/2$ pairs of the marker sets is measured by means of a similarity index. Fourth, the reproducibility performance of the algorithm (C_{avg}) is expressed as the mean of the all pairwise similarities, i.e.,

$$C_{avg} = \frac{2\sum_{i=1}^{K}\sum_{j=i+1}^{K} SI(\mathcal{F}_i, \mathcal{F}_j)}{K(K-1)}, \tag{28}$$

where SI stands for the similarity index that measures the similarity between any two marker sets $\mathcal{F}_i$ and $\mathcal{F}_j$. Among the variety of similarity indexes that have been proposed, the Kuncheva index (KI) [41] was adopted as a measure of similarity in this work. This is because KI includes a correction term to account for the possible bias that results from the existence of common markers among the two signature lists that are randomly selected. Formally, KI is expressed as:

$$KI(\mathcal{F}_i, \mathcal{F}_j) = \frac{p.|\mathcal{F}_i \cap \mathcal{F}_j| - |\mathcal{F}|^2}{|\mathcal{F}|\,(p - |\mathcal{F}|)} = \frac{|\mathcal{F}_i \cap \mathcal{F}_j| - (|\mathcal{F}|^2/p)}{|\mathcal{F}| - (|\mathcal{F}|^2/p)} \tag{29}$$

where $|\mathcal{F}|$ represents the size of the identified markers (i.e., $|\mathcal{F}| = |\mathcal{F}_i| = |\mathcal{F}_j|$). The values of Kuncheva index range from -1 to 1. Larger KI values indicate higher stability performance. Due to the correction term ($|\mathcal{F}|^{2/p}$), which accounts for selecting markers that are common among marker sets due to chance, the KI may take negative values.

In this paper, the stability performance was visualized by presenting three types of descriptive plots. The first plot shows the average KI over all pairwise comparisons.

The second plot provides more details about the distribution of all the KI values by presenting their histogram. An ideal algorithm in terms of stability will have the Dirac-delta distribution at KI equal to 1. This means that the algorithm generates the same set of markers over all subsamples. Practically, the more concentrated the histogram is to the right side of the plot, the more stable is the algorithm. The third plot aims to depict the stability of the ranked microbial marker lists. This is achieved by ordering all the selected markers based on their ranks. Then, a boxplot is generated for the ranks obtained in all the K subsamples for each selected marker. A perfect algorithm in the sense of stability of the ranked lists will have boxplots that are centered at the 45° line, which means that the algorithm perfectly preserves the order of the detected markers in all subsamples.

Classification performance

Accuracy, sensitivity, and specificity are the three metrics that were used to measure the classification performance. The classification accuracy represents the fraction of the number of samples that were correctly predicted to the total number of samples. One major drawback of accuracy is that its value is dominated by the class with the majority of samples. Therefore, in case of imbalanced class distribution or when the forecast of the minority group is critical, accuracy may be misleading. Thus, class-specific measures (i.e., sensitivity and specificity) are needed to provide a more accurate picture about the classification performance. Sensitivity (specificity) is expressed as the contribution of the correct predictions in the positive (negative) class. Formally, let TN and TP represent the number of correctly identified negative and positive subjects. Consider that FN and FP represent the number of false-predicted instances in the negative and positive classes, respectively. The accuracy, sensitivity and specificity measures are expressed as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{30}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{31}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{32}$$

The classification performance is measured empirically according to the evaluation protocol shown in Algorithm 2. At the k^{th} iteration, the classifier is trained by the data corresponding to the selected markers $\left(\mathbf{D}_k^{train}(\mathcal{F}_k)\right)$. Then it is tested against the remaining $\mathbf{D}_k^{test}(\mathcal{F}_k)$. One major benefit from repeating the evaluation K times is to mitigate the over-optimistic results that are associated with the conventional cross-validation on small-sample studies [42]. In our experiments, two versions of the nearest centroid classifiers were employed. The first version relies on the l_1 norm, while the second version exploits the l_2 norm. Therefore, herein paper, the first classifier is referred to as NCC-1, while the second one is denoted as NCC-2.

Discussion of evaluation criteria

A critical challenge for assessing the performnce of biomarker detection approaches is the lack of information about the true biomarkers. This hampers the objective assessment of the performance of competing biomarker selection algorithms. To overcome this challenge, evaluation criteria have to be properly developed to replicate comparisons as if the true markers were known. The evaluation criteria have to capture the features of the true biomarkers. The true biomarkers exhibit two properties. The first feature is the fact that the true biomarkers must allow differentiating different phenotypes. In general, this is assessed via the performance of a classifier designed based on the selected biomarkers. The second feature relies on the fact that true signatures appear not to be sensitive against variations in the training samples. This feature is evaluated via empirical assessment of the biomarker identification algorithm stability.

A common practice is to use *only* the classification performance as a measure of the effectiveness of a biomarker detection algorithm. In addition to ignoring the reproducibility performance, relying solely on the classification performance may be misleading for several reasons. First, the classification performance depends on factors other than the quality of the selected variables (i.e., biomarkers). In particular, the preprocessing steps and classifier model employed significantly impact the classification performance. Second, in the small sample size setups, the empirical estimation of classification accuracy may not reflect the true performance of a classifier.

Unfortunately, the existing metagenomic biomarker identification schemes have not yet considered the reproducibility performance in their assessments. This calls the utility of these methods under question. Similarly, assessing a biomarker detection algorithm based on its stability performance is delusive. For example, a trivial algorithm that returns the same features irrespective of the training samples will achieve a perfect stability performance. Thus, reproducibility needs to be assessed together with the classification performance.

Simulation setup

The classification and consistency metrics were used to measure the efficiency of the six biomarker detection algorithms in identifying potential markers. The consistency-classification evaluation protocol is presented in Algorithm 2. In our studies, a random subsampling without replacement is utilized to generate 500 subsamples (i.e., $K = 500$) variations of the original dataset. Each subsample contains 80% of the samples in

the original dataset (i.e., $r = 0.8$). The classification and consistency performance were evaluated at different number of selected markers to provide further insights on the performance of the competing algorithms under varying sizes of the biomarker sets. The reported outcomes stand for the average over the 500 experiments.

The classification performance is measured empirically according to the evaluation protocol shown in Algorithm 2. At the k^{th} iteration, the classifier is trained by the data corresponding to the selected markers $\left(\mathbf{D}_k^{train}(\mathcal{F}_k)\right)$. Then it is tested against the remaining $\mathbf{D}_k^{test}(\mathcal{F}_k)$. One major benefit from repeating the evaluation K times is to mitigate the over-optimistic results that are associated with the conventional cross-validation on small-sample studies [42]. In our experiments, two variants of the nearest centroid classifiers were used. The first approach employed the l_1 norm as a measure of distance, while in the second approach, the l_2 norm was used. In this paper, we refer to the first classifier as NCC-1 and to the second one as NCC-2.

Discussion of evaluation criteria

A major bottleneck for the evaluation of biomarker discovery algorithms is the lack of knowledge of the true biomarkers. This hampers the objective assessment of the performance of competing biomarker selection algorithms. To overcome this challenge, evaluation criteria have to be suitably designed in order to mimic comparisons as if the true markers were known. In particular, the evaluation metrics need to capture the features of the true biomarkers. True biomarkers are characterized by two properties. The first property is that the true markers enable distinguishing between different phenotypes. Commonly, this feature is measured via the classification performance of a classifier model built using only the selected biomarkers. The second feature is that true signatures tend to be robust against the variation in the training set. This feature can be assessed through empirical estimation of the stability of the biomarker detection algorithm.

A common practice is to use *only* the classification performance as a measure of the effectiveness of a biomarker detection algorithm. In addition to ignoring the reproducibility performance, relying solely on the classification performance may be misleading for several reasons. First, the classification performance depends on factors other than the quality of the selected variables (i.e., biomarkers). In particular, the preprocessing steps and classifier model employed significantly impact the classification performance. Second, in the small sample size setups, the empirical estimation of classification accuracy may not reflect the true performance of a classifier.

Surprisingly, the existing state-of-art metagenomic biomarker detection algorithms have not considered the reproducibility performance in their assessment. This calls the utility of these methods under question. Similarly, assessing a biomarker detection algorithm based on its stability performance is delusive. For example, a trivial algorithm that returns the same features irrespective of the training samples will achieve a perfect stability performance. Thus, reproducibility needs to be assessed together with the classification performance.

Simulation setup

The classification and consistency metrics were used to measure the efficiency of the six biomarker detection algorithms in identifying potential markers. The consistency-classification evaluation protocol is shown in Algorithm 2. In our experiments, a random subsampling without replacement is utilized to generate 500 subsamples (i.e., $K = 500$) variations of the original dataset. Each subsample contains 80% of the samples in the original dataset (i.e., $r = 0.8$). The classification and consistency performance were evaluated at different number of selected markers to provide further insights on the performance of the competing algorithms under varying sizes of the biomarker sets. The reported results represent the average over the 500 experiments.

Dogs with exocrine pancreatic insufficiency (EPI) dataset

The reproducibility performance in terms of the average KI stability values over all the pairwise comparisons (i.e., $K(K-1)/2 = 124750$ comparisons; $K = 500$) of the six algorithms for a changing number of biomarkers from the EPI dataset is illustrated in Fig. 1. As it is transparent from Fig. 1, RegLRSD outperforms all the other algorithms. The improvement gain of RegLRSD over the other algorithms

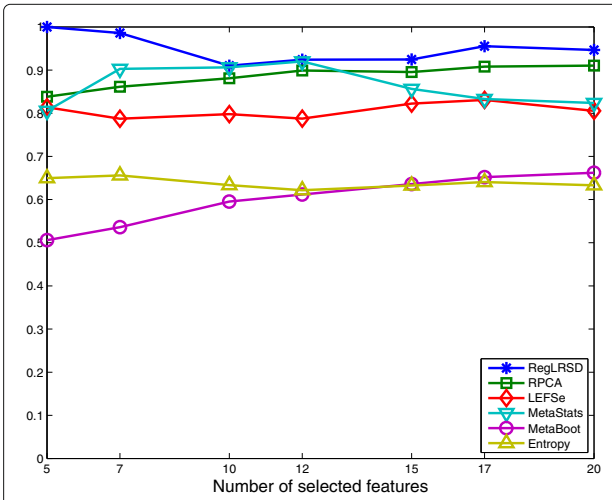

Fig. 1 Average of Kuncheva Index (KI) at varying number of selected markers for the six biomarker detection algorithms over the dogs with EPI dataset

in terms of reproducibility performance is higher at lower number of selected markers. This indicates that RegLRSD is more certain in identifying small subsets of potential markers.

Figure 2 presents the histogram of the KI index computed over the 124750 pairwise comparisons when the size of the selected biomarkers equals 20. The concentration of the histogram of RegLRSD at high KI values reveals that the RegLRSD algorithm achieves a high reproducibility performance. In particular, RegLRSD provides a stability value that is larger than or equal to 90% for almost 90% of the times. On the other hand, the other algorithms are less prone to achieve the same stability performance. In particular, RPCA, LEFSe, and MetaStats yield a stability performance that is larger than or equal to 90% for only 75, 15, and 30% of the times, respectively, and less than 5% of the times for both MetaBoot, and entropy-based algorithm. Moreover, the spread of the histograms of LEFSe, MetaStats, MetaBoot and entropy algorithms over wide range of KI values indicates a serious inconsistency problem that puts the outcomes of these algorithms under question.

The ranking stability of the selected microbial signatures over all the $K = 500$ variations of the original dataset is depicted in Fig. 3. In addition to the high reproducibility performance, the RegLRSD algorithm corroborates its ability to preserve the order (i.e., rank) of the selected markers as revealed from the concentration of the boxplots of the ranks around the 45° line. The spread of the rank boxplots of the other algorithms indicates that the rank of the selected markers in these algorithms varies significantly with respect to small variations in the dataset.

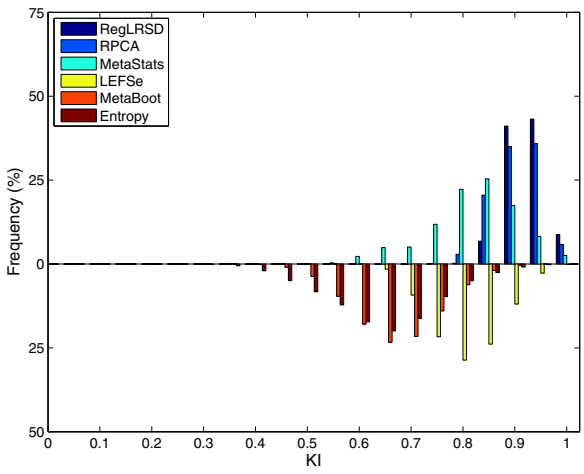

Fig. 2 Histogram plots of the KI values generated by the six biomarker detection algorithms over the dogs with EPI dataset. Each histogram is created using 124750 values of KI which are generated from all pairwise comparisons over the $K = 500$ runs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons)

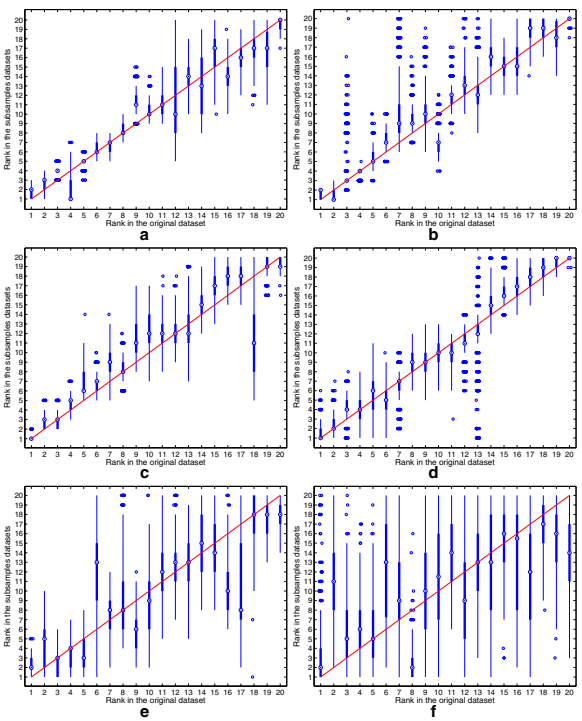

Fig. 3 Rank boxplots in the subsamples against rank in the original data set for the six algorithms over the dogs with EPI dataset. **a** RegLRSD. **b** RPCA. **c** LEFSe. **d** MetaStats. **e** MetaBoot. **f** Entropy

For example, the rank of the marker that is ranked sixth when applying the MetaBoot algorithm over the original dataset varies significantly over 500 different subsamples as cleared from Fig. 3.e. Specifically, the median value for all these ranks (i.e., ranks obtained in the 500 subsamples) equals 13 and the interquartile range (IQR) equals 6 (from 9 to 15). Moreover, in some subsamples, this marker was ranked first, while in other subsamples it was ranked twentieth.

The classification performance of the competing algorithms is illustrated in Fig. 4. The first column in Fig. 4 depicts the outcomes for the NCC-1 classifier, while the second column illustrates the outcomes for the NCC-2 classifier. In general, all the algorithms yield a robust performance regardless of the number of selected biomarkers. The identified markers by RegLRSD, LEFSe, MetaStats, and MetaBoot show high ability to distinguish between healthy and diseased samples related to EPI as revealed by the high accuracy, sensitivity and specificity of these algorithms compared to RPCA and entropy algorithms, especially when the NCC-2 is used. The better performance of RegLRSD compared to RPCA demonstrates that incorporating the prior knowledge improves the performance markedly.

Figure 5 displays the top 20 identified markers by RegLRSD and their scores. RegLRSD suggests that the

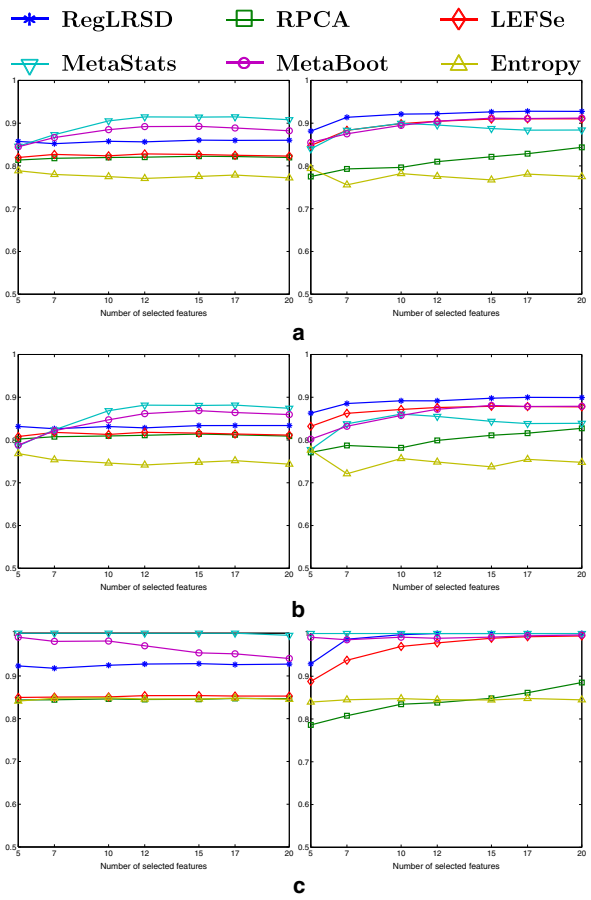

Fig. 4 Classification performance of the six biomarker detection algorithms using the two classification algorithms, NCC-1 (column 1) and NCC-2 (column 2), over the dogs with EPI dataset. Classification performance is measured in terms of **a** accuracy, **b** sensitivity and **c** specificity

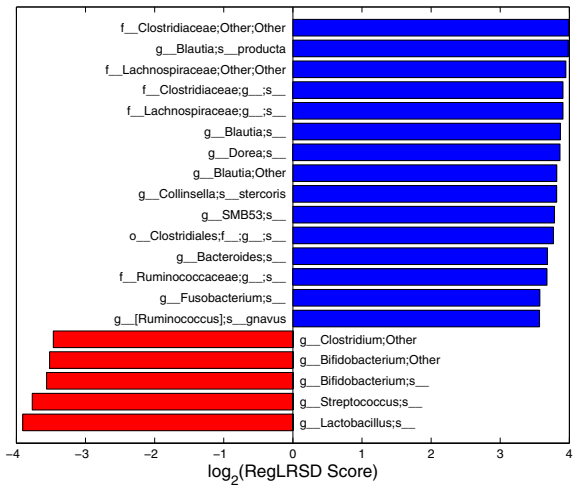

Fig. 5 Biomarker discovery results when we selected the top 20 markers from the dogs with EPI dataset using the RegLRSD algorithm. *Blue* and *red* bars represent microbes that are enriched in control and diseased samples, respectively

EPI may be characterized by the decrease in Blautia, Bacteroides, Fusobacterium, Ruminococcus genera in dogs with EPI. Also, the genera, Lactobacillus, Streptococcus, Bifidobacterium present a significant growth in their abundance levels in dogs with EPI when compared to healthy dogs. Previous studies have also showed an increase in Lactobacillus and Streptococcus abundance levels in dogs with EPI. In particular, two culture-based investigations reported an increased number of Lactobacillus and Streptococcus in the duodenum [43], jejunum and colon of dogs with EPI [44].

Dogs with idiopathic inflammatory bowel disease (IBD) dataset

The stability performance measured in terms of the average KI values for the six algorithms over different numbers of biomarkers is depicted in Fig. 6. The results in Fig. 6 illustrate that RegLRSD outperforms the rest of the algorithms in terms of reproducibility performance. Moreover, adding the smoothing constraint in RegLRSD results in an improvement in the stability performance by almost $2-7\%$ over the standard RPCA. Noticeably, LEFSe and MetaBoot provide a poor reproducibility performance. For example, the average KI values range around $30\%-50\%$ for MetaBoot and around $40\%-65\%$ for LEFSe.

The histograms of the KI index computed over the 124750 pairwise comparisons when the size of the selected biomarkers equals 20 is depicted in Fig. 7. The histogram of RegLRSD illustrates the superior

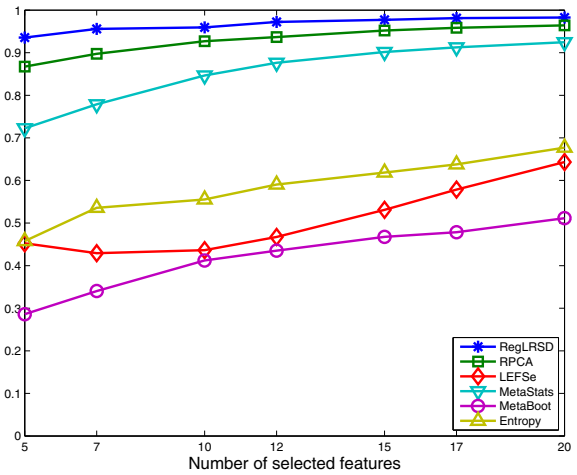

Fig. 6 Average of Kuncheva Index (KI) at varying number of selected markers for the six biomarker detection algorithms over the dogs with IBD dataset

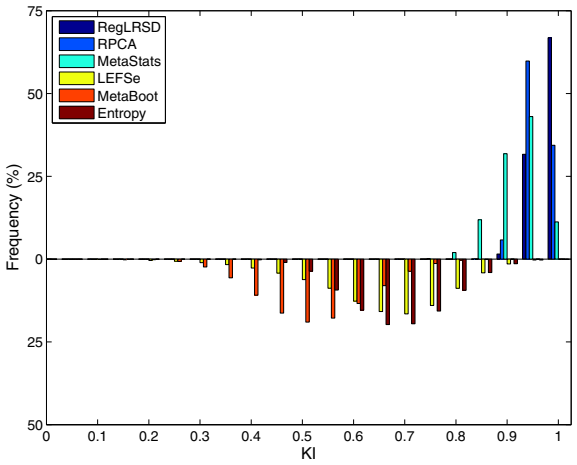

Fig. 7 Histogram plots of the KI values generated by the six biomarker detection algorithms over the dogs with IBD dataset. Each histogram is created using 124750 values of KI which are generated from all pairwise comparisons over the $K = 500$ runs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons)

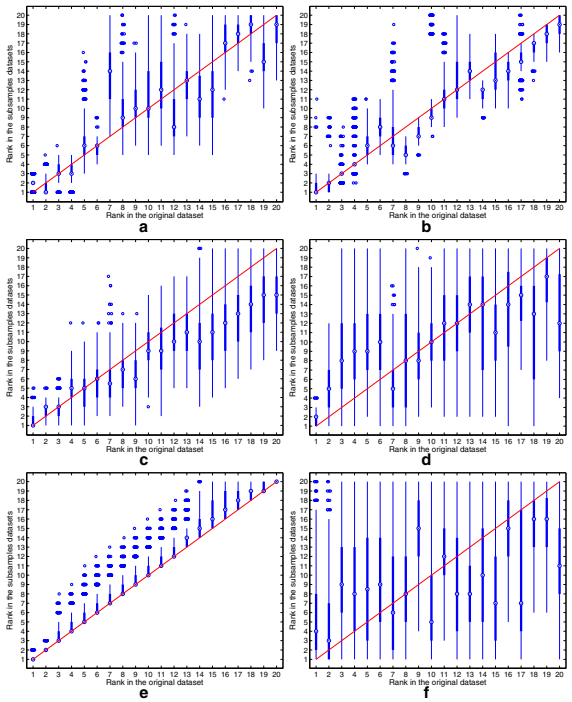

Fig. 8 Rank boxplots in the subsamples against rank in the original data set for the six algorithms over the dogs with IBD dataset. **a** RegLRSD. **b** RPCA. **c** LEFSe. **d** MetaStats. **e** MetaBoot. **f** Entropy

performance of RegLRSD as it achieves 100% stability for more than 65% of the times. RPCA and MetaStats show an adequate consistency. On the other hand, LEFSe, MetaBoot, and entropy-based approach tend to provide poor performance as their corresponding histograms are centered at low KI values and spread over wide range of KI values.

The ranking stability of the selected microbial signatures over all the $K = 500$ subsamples is presented in Fig. 8. The rank of the selected markers by RegLRSD, RPCA, and MetaBoot is more consistent against the variation in the dataset. This contrasts the performance of the LEFSe, MetaStats, and entropy-based algorithms, in which the importance (i.e., rank) of the selected features varies drastically due to adding/removing a small number of samples from the original dataset. In terms of classification performance, the RegLRSD algorithm outperforms the other algorithms especially when the NCC-2 classifier is used as revealed from Fig. 9. Noticeably, RegLRSD yields a significant improvement over the RPCA algorithm. This reflects the efficiency of incorporating the prior knowledge information in generating more accurate results.

RegLRSD suggested several bacterial groups as potential markers for IBD. The top 20 detected biomarkers by the RegLRSD algorithm and their scores are displayed in Fig. 10. At higher phylogenetic levels, the majority of these bacterial groups belong to Firmicutes, Bacteroidetes, and Proteobacteria. In particular, the Enterobacteriaceae is the main driver for increasing the abundance level of Gammaproteobacteria in dogs with IBD. The quantitative PCR (qPCR) assays suggest that this increase is mainly due to Escherichia coli (i.e., E. coli). Several studies in human patients with IBD [45, 46] reported that E. Coli exhibits virulent potential such as adhesive capacity, invasive capacity, toxin production, and inflammatory cytokine stimulation. Similarly, the results in [47] associated several adherent and invasive strains of E. Coli with granulomatous colitis in boxer dogs. RegLRSD have suggested several genera belonging to Firmicutes to be as a potential markers for IBD. In particular, Blautia, Turicibacter, and Faecalibacterium were decreased in IBD. Most of these bacterial groups belong to Clostridium clusters IV and XIVa and are recognized as the major producer of several metabolites including short chain fatty acids (SCFA). Consequently, decreasing the abundance level of these bacterial groups may impact the host health. These findings comply with previous studies in duodenal mucosal/luminal content and feces in dogs with IBD [48–50].

Mouse model of ulcerative colitis (UC) dataset

The mean KI values across all the pairwise comparisons and their histograms in the presence of the 20 selected biomarkers are presented in Figs. 11 and 12, respectively. Figure 11 demonstrates that RegLRSD outperforms all the other algorithms and exhibits a high reproducibility

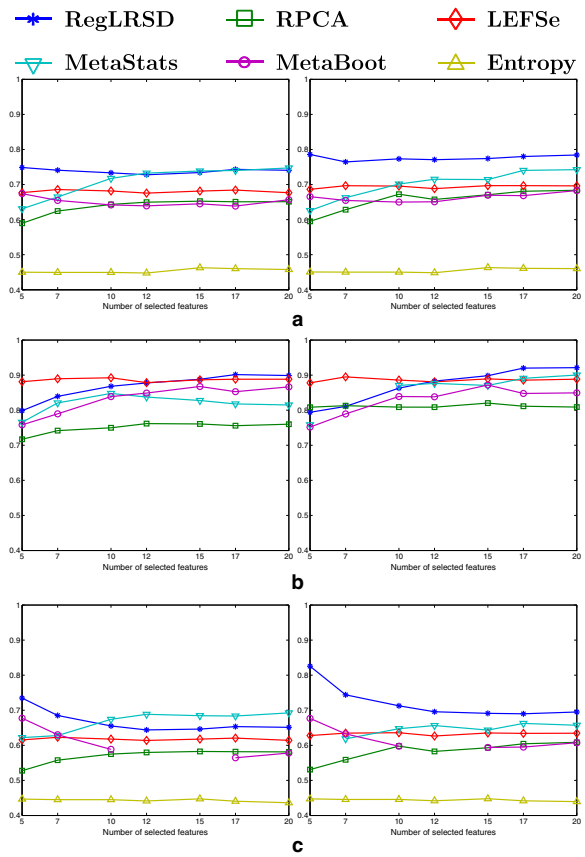

Fig. 9 Classification performance of the six biomarker detection algorithms using the two classification algorithms, NCC-1 (column 1) and NCC-2 (column 2), over the dogs with IBD dataset. Classification performance is measured in terms of **a** accuracy, **b** sensitivity and **c** specificity

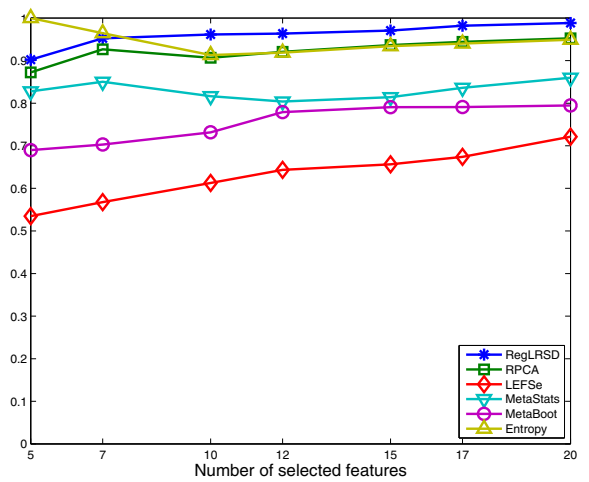

Fig. 11 Average of Kuncheva Index (KI) at varying number of selected markers for the six biomarker detection algorithms over the mouse model of UC dataset

performance. In particular, the improvement gain is about 5% over RPCA and entropy-based algorithm, 15% over MetaStats, 20 − 25% over MetaBoot, and more than 30% over LEFSe.

The ranking stability of the selected microbial signatures over all the $K = 500$ subsamples is illustrated in Fig. 13. The outcomes in Fig. 13 point a serious inconsistency problem in the performance of LEFSe, MetaStats and entropy-based algorithm. The two matrix decomposition-based algorithms (i.e., RegLRSD and RPCA) provide a comparable performance in terms

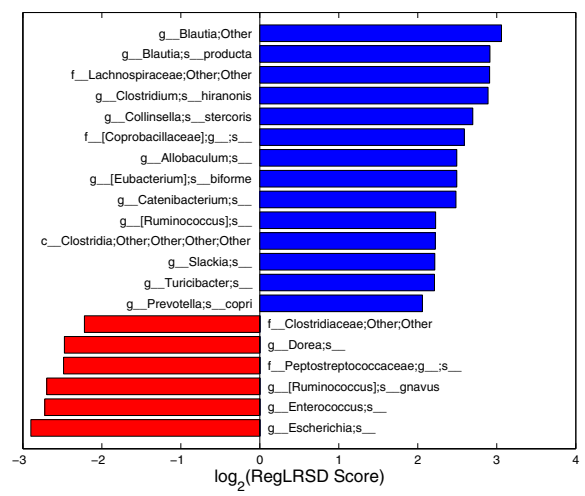

Fig. 10 Biomarker discovery results when we selected the top 20 markers from the dogs with IBD dataset using the RegLRSD algorithm. *Blue* and *red* bars represent microbes that are enriched in control and diseased samples, respectively

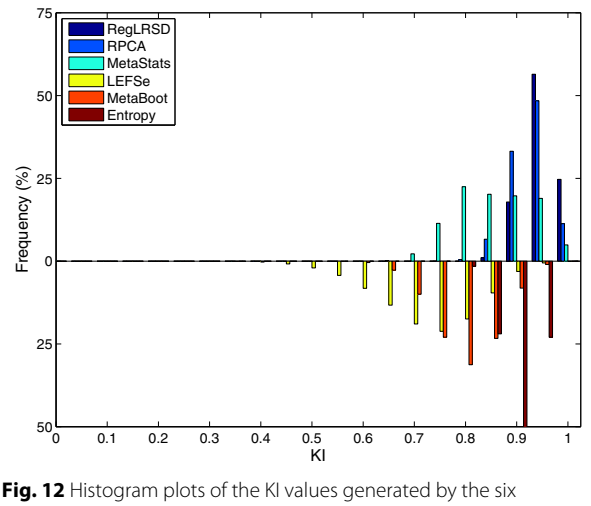

Fig. 12 Histogram plots of the KI values generated by the six biomarker detection algorithms over the mouse model of UC dataset. Each histogram is created using 124750 values of KI which are generated from all pairwise comparisons over the $K = 500$ runs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons)

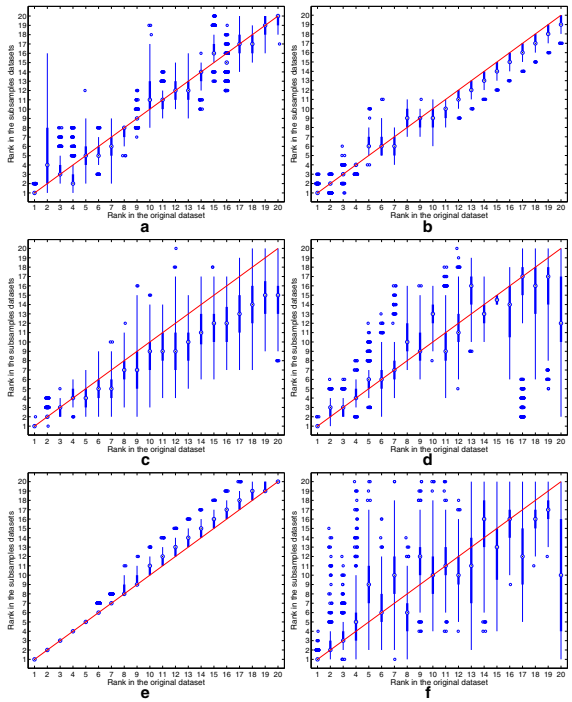

Fig. 13 Rank boxplots in the subsamples against the rank in the original data set for the six algorithms over the mouse model of UC dataset.
a RegLRSD. **b** RPCA. **c** LEFSe. **d** MetaStats. **e** MetaBoot. **f** Entropy

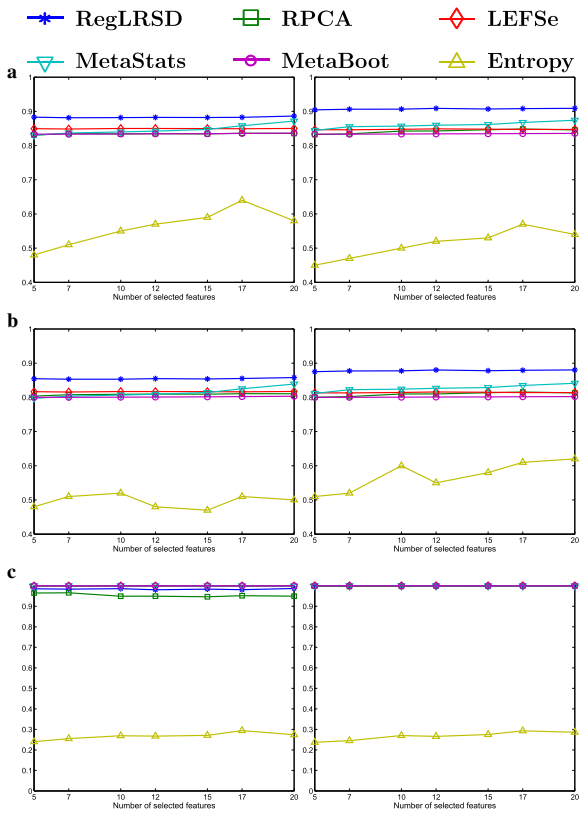

Fig. 14 Classification performance of the six biomarker detection algorithms using the two classification algorithms, NCC-1 (column 1) and NCC-2 (column 2), over the mouse model of UC dataset. Classification performance is measured in terms of **a** accuracy, **b** sensitivity and **c** specificity

of retaining the rank of the selected markers over different subsamples of the data set.

The classification performance of the six algorithms in the presence of a changing number of biomarkers from the UC mice model data set is illustrated in Fig. 14. The results in Fig. 14 point out that all the algorithms, except the entropy-based algorithm, provide almost the same classification accuracy (i.e., 80 − 84%).

The top 15 detected biomarkers by the RegLRSD algorithm are depicted in Fig. 15. The majority of these markers comply with the previous studies. For example, the authors of [51, 52] reported reduced concentrations of Lactobacillus and Bifidobacterium in colonic biopsy specimens in patients with active UC. The study [53] has suggested that the UC could be depicted via a decline in the abundance levels of Bacteroides. The authors of [9] reported that the decrease in the abundance levels of acetate producer clades such as Ruminococcaceae may reduce the host capability to fix the epithelium and to regulate inflammation. This may explain the selection of Oscillibacter, which belongs to Ruminococcaceae, as possible marker for UC. Subjects with UC showed significant reduction in Helicobacter pylori [54], the most well-known known species of Helicobacter genus.

Conclusions

Recent advancements in metagenomic sequencing associated microbes with several health and disease states of the host. Identifying potential metagenomic markers is essential for understanding biological systems

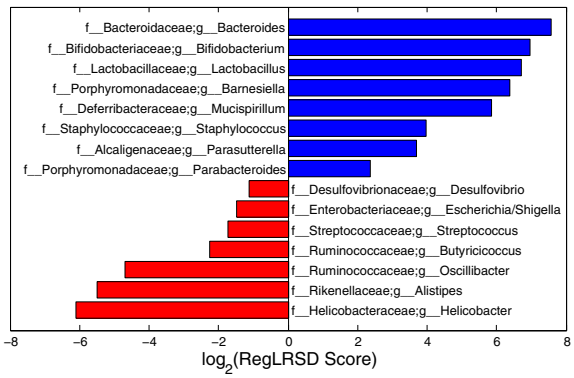

Fig. 15 Biomarker discovery results when we selected the top 15 markers from the mouse model of UC dataset using the RegLRSD algorithm. *Blue* and *red* bars represent microbes that are enriched in control and diseased samples, respectively

and designing possible therapies for diseases. Therefore, developing robust and stable biomarker detection algorithms is crucial in order to infer correct biological statements and translate these results into clinical practice. Herein paper, we developed the RegLRSD algorithm for biomarker detection. Apart from the conventional statistical and feature selection frameworks to tackle the problem of finding potential metagenomic biomarkers, RegLRSD formulates the biomarker detection as a matrix decomposition problem. In particular, RegLRSD models the abundance profiles of relevant and irrelevant microbes as sparse and low-rank matrices, respectively. This renders identifying potential biomarkers as the problem of decomposing the bacterial abundance data matrix into a sparse matrix and a low-rank matrix.

To enhance the accuracy of estimating the low-rank matrix and the sparse matrix, RegLRSD constrains the low rank matrix to be smooth in order to integrate the prior knowledge that the abundance profiles of irrelevant bacteria do not exhibit strong variation between different phenotypes in the biomarker detection process. Then we developed an efficient solution for this decomposition problem by exploting the alternating direction method of multipliers. In addition to the computationally efficient solution for RegLRSD, a major advantage of RegLRSD is the convex formulation of the biomarker detection problem. This convex formulation enables adding convex constraints that reflect our prior knowledge about the biological system under study. These additional constraints help in designing better algorithms that are more accurate and provide more consistent biological findings. The improved performance of RegLRSD over the conventional RPCA algorithm (i.e., without the smoothness constraint) demonstrates the efficiency of incorporating prior knowledge in the design of a biomarker detection algorithm.

In addition to the development of a novel algorithm for identifying metagenomic markers (i.e., RegLRSD), this paper addressed an important feature of the metagenomic biomarker discovery algorithms. This feature is the ability of biomarker detection algorithms to generate reproducible results. This is crucial to translate the outcome of these algorithms into practical applications. Surprisingly, the stability/reproducibility performance was not addressed by the existing metagenomic biomarker identification algorithms. Our simulation results demonstrate that the existing methods for metagenomic biomarker discovery present poor reproducibility performance. In particular, the spread of the histograms of LEFSe, MetaStats, MetaBoot and entropy-based algorithm over a wide range of KI values indicates a serious inconsistency problem that puts the outcomes of these algorithms under question.

Comprehensive comparisons with the latest biomarker detection approaches were conducted. In particular, RegLRSD was contrasted with two statistical-based approaches (i.e., LEFSe and MetaStats), two machine learning-based algorithms (MetaBoot and entropy) and a reduced from of RegLRSD in which the smoothness constraint is not considered (i.e., RPCA). The competing algorithms were tested against three realistic metagenomic datasets. The first and second datasets pertain to healthy dogs and dogs diagnosed with EPI and IBD, respectively. The third dataset refers to a mouse model of UC. These approaches were assessed in terms of classification accuracy and reproducibility performance. The simulation results show that the detected markers by RegLRSD enable discriminating metagenomic samples belonging to different phenotypes with a quite high accuracy. Moreover, RegLRSD exhibits superior consistency performance when compared to other algorithms. This renders the RegLRSD algorithm as a robust and reliable tool to identify potential metagenomic markers that may characterize the difference between samples belonging to different phenotypes.

The results presented in this paper demonstrate that the two matrix decomposition-based algorithms (i.e., RegLRSD and RPCA) are successful in providing high reproducibility and classification accuracy performance compared to the conventional statistical and machine learning-based algorithms. This validates the idea of modeling the bacterial abundance data matrix as the superposition of a low-rank matrix representing the uninformative microbes and a sparse matrix containing the abundances of informative microbes. Moreover, the improvement in the performance of RegLRSD compared to RPCA demonstrates (i) the validity of our assumption that the abundance profiles of irrelevant bacteria are smooth, and (ii) incorporating prior knowledge in the design of a biomarker detection algorithm may lead to more robust results.

Due to the necessity of developing user-friendly tools that enable the researchers to analyze metagenomic data, RegLRSD is implemented as a standalone executable software package and is made available at https://sites.google.com/a/tamu.edu/mustafa/software/reglrsd.

Abbreviations

ADMM: Alternating direction method of multipliers; EPI: Exocrine pancreatic insufficiency; IBD: Inflammatory bowel disease; KI: Kuncheva index; LDA: Linear discriminant analysis; NCBI: National center for biotechnology information NCC: Nearest centroid classifier; OTU: Operational taxonomic unit; PCA: Principal component analysis; PCP: Principal component pursuit; PLS: Partial least square; QIIME: Quantitative insights into microbial ecology; RDP: Ribosomal database project; RegLRSD: Regularized low rank-sparse decomposition; RPCA: Robust principal component analysis; UC: Ulcerative colitis

Acknowledgements

Not applicable.

Funding
This work was supported by the German Jordanian University and a Proof-of-Concept Grant offered by Texas A&M University in College Station. The funding body did not play any role in the design or conclusion of your study.

Availability of data and materials
A standalone executable software packages of the RegLRSD algorithm for both Windows and Linux systems are available at https://sites.google.com/a/tamu.edu/mustafa/software/reglrsd
The sequencing data for the dogs with EPI were deposited into the NCBI-SRA database under the accession number SRP091334 and the hyperlink to the dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRP040310
The sequencing data for the dogs with IBD were deposited into the NCBI-SRA database under the accession number SRP040310 and the hyperlink to the dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRP091334.
The mouse model of ulcerative colitis dataset can be found in the supplementary material of [13].

Authors' contributions
MA conceived of the study, developed the framework, conducted the analysis, provided the results interpretation and wrote the manuscript. AB contributed to the framework development and analysis, and helped in reviewing the manuscript. ES and JS provided an overall guidance, participated in the statistical and biological interpretation of the results, and were involved in drafting the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Bioinformatics and Genomic Signal Processing Lab, ECEN Dept., Texas A&M University, 77843-3128, College Station, TX, USA. [2]College of Veterinary Medicine and Biomedical Sciences, Gastrointestinal Laboratory, Texas A&M University, 77843-3128, College Station, TX, USA.

Received: 24 April 2017 Accepted: 22 June 2017
Published online: 10 July 2017

References

1. Schloss PD, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 2005;6(8):229.
2. Jurkowski A, Reid AH, Labov JB. Metagenomics: a call for bringing a new science into the classroom (while it's still new). CBE-Life Sci Educ. 2007;6(4):260–5.
3. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.
4. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. Nature. 2008;452(7187):629–32.
5. Flint HJ. Obesity and the gut microbiota. J Clin Gastroenterol. 2011;45: S128—32.
6. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.
7. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science. 2013;341(6150):241214.
8. Larsen N, Vogensen FK, Van Den Berg F, Nielsen DS, Andreasen AS, Pedersen BK, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PloS One. 2010;5(2):e9085.
9. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79.
10. Moore W, Moore LH. Intestinal floras of populations that have a high risk of colon cancer. Appl Environ Microbiol. 1995;61(9):3202–7.
11. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, et al. Human gut microbiome and risk of colorectal cancer. J Natl Cancer Inst. 2013: djt300.
12. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009;5(4):e1000352.
13. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.
14. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics. 2003;19(12):1484–91.
15. Simon R. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). ACM SIGKDD Explor Newslett. 2003;5(2):31–6.
16. Wooley JC, Ye Y. Metagenomics: facts and artifacts, and computational challenges. J Comput Sci Technol. 2010;25(1):71–81.
17. Swan KA, Curtis DE, McKusick KB, Voinov AV, Mapa FA, Cancilla MR. High-throughput gene mapping in Caenorhabditis elegans. Genome Res. 2002;12(7):1100–5.
18. Khosravi A, Mazmanian SK. Disruption of the gut microbiome as a risk factor for microbial infections. Curr Opin Microbiol. 2013;16(2): 221–7.
19. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
20. Bucci V, Nadell CD, Xavier JB. The evolution of bacteriocin production in bacterial biofilms. The Am Nat. 2011;178(6):E162—73.
21. Klitgord N, Segre D. Environments that induce synthetic microbial ecosystems. PLoS Comput Biol. 2010;6(11):e1001002.
22. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3(1):140.
23. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.
24. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005;365(9460):671–9.
25. Wright J, Ganesh A, Rao S, Peng Y, Ma Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in Neural Information Processing Systems. New York: Curran Associates, Inc.; 2009. p. 2080–8.
26. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? J ACM (JACM). 2011;58(3):11.
27. Donoho DL. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. Commun Pur Appl Math. 2006;59(6):797–829.
28. Recht B, Fazel M, Parrilo PA. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. 2010;52(3):471–501.
29. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20(4):1956–82.
30. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenom. 1992;60(1):259–68.
31. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends® in Mach Learn. 2011;3(1):1–122.
32. Liu J, Yuan L, Ye J. An efficient algorithm for a class of fused lasso problems. In: Proceedings of the 16th, ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2010. p. 323–32.
33. Ghadimi E, Teixeira A, Shames I, Johansson M. Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems. IEEE Trans Autom Control. 2015;60(3):644–58.
34. Liu J, Ji S, Ye J. SLEP: Sparse Learning with Efficient Projections. 2009. Available from http://www.public.asu.edu/jye02/Software/SLEP. Accessed 12 Sept 2016.

35. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
36. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72.
37. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
38. Veiga P, Gallini CA, Beal C, Michaud M, Delaney ML, DuBois A, et al. Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes. Proc Natl Acad Sci. 2010;107(42):18132–7.
39. Wang X, Su X, Cui X, Ning K. MetaBoot: a machine learning framework of taxonomical biomarker discovery for different microbial communities based on metagenomic data. PeerJ. 2015;3:e993.
40. Alshawaqfeh M, Bashaireh A, Serpedin E, Suchodolski J. Consistent metagenomic biomarker detection via robust PCA. Biol Direct. 2017;12(1):4.
41. Kuncheva LI. A stability index for feature selection. In: Artificial Intelligence and Applications. Anaheim, CA: ACTA Press; 2007. p. 421–7.
42. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004;20(3):374–80.
43. Westermarck E, Myllys V, Aho M. Effect of treatment on the jejunal and colonic bacterial flora of dogs with exocrine pancreatic insufficiency. Pancreas. 1993;8(5):559–62.
44. Simpson K, Batt R, Jones D, Morton D. Effects of exocrine pancreatic insufficiency and replacement therapy on the bacterial flora of the duodenum in dogs. Am J Vet Res. 1990;51(2):203–6.
45. Kotlowski R, Bernstein CN, Sepehri S, Krause DO. High prevalence of Escherichia coli belonging to the B2+ D phylogenetic group in inflammatory bowel disease. Gut. 2007;56(5):669–75.
46. De la Fuente M, Franchi L, Araya D, Díaz-Jiménez D, Olivares M, Álvarez-Lobos M, et al. Escherichia coli isolates from inflammatory bowel diseases patients survive in macrophages and activate NLRP3 inflammasome. Int J Med Microbiol. 2014;304(3):384–92.
47. Simpson KW, Dogan B, Rishniw M, Goldstein RE, Klaessig S, McDonough PL, et al. Adherent and invasive Escherichia coli is associated with granulomatous colitis in boxer dogs. Infect Immun. 2006;74(8):4778–92.
48. Suchodolski JS, Camacho J, Steiner JM. Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16S rRNA gene analysis. FEMS Microbiol Ecol. 2008;66(3):567–78.
49. Suchodolski JS, Xenoulis PG, Paddock CG, Steiner JM, Jergens AE. Molecular analysis of the bacterial microbiota in duodenal biopsies from dogs with idiopathic inflammatory bowel disease. Vet Microbiol. 2010;142(3):394–400.
50. Rossi G, Pengo G, Caldin M, Piccionello AP, Steiner JM, Cohen ND, et al. Comparison of microbiological, histological, and immunomodulatory parameters in response to treatment with either combination therapy with prednisone and metronidazole or probiotic VSL# 3 strains in dogs with idiopathic inflammatory bowel disease. PloS One. 2014;9(4):e94699.
51. Ruseler-van Embden J, Schouten W, Van Lieshout L. Pouchitis: result of microbial imbalance? Gut. 1994;35(5):658–64.
52. Poxton I, Brown R, Sawyerr A, Ferguson A. Mucosa-associated bacterial flora of the human colon. J Med Microbiol. 1997;46(1):85–91.
53. Sasaki M, Klapproth JMA. The role of bacteria in the pathogenesis of ulcerative colitis. J Signal Transduct. 2012;2012.
54. Jin X, Chen Y, Chen S, Xiang Z. Association between Helicobacter Pylori infection and ulcerative colitis–a case control study from China. Int J Med Sci. 2013;10(11):1479–84.