Infer disease-associated microbial biomarkers based on metagenomic and metatranscriptomic data

Zhaoqian Liu[1], Qi Wang[1], Dongjun Chung[2], Qin Ma[2], Jing Zhao[2,#], Bingqiang Liu[1,#]

[1] School of Mathematics, Shandong University, Jinan 250200, China.

[2] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA.

[#]To whom correspondence should be addressed.

Email

Zhaoqian Liu --- lzqian8023@163.com;

Qi Wang --- wangqi1994_sdu@163.com;

Dongjun Chung --- dongjun.chung@osumc.edu;

Qin Ma --- qin.ma@osumc.edu;

Jing Zhao --- jing.zhao2@osumc.edu;

Bingqiang Liu --- bingqiang@sdu.edu.cn.

Abstract

Unveiling disease-associated microbial biomarkers (e.g., key species, genes, and pathways) is an efficient strategy for the diagnosis and therapy of diseases. However, the heterogeneity and large size of microbial data bring tremendous challenges for fundamental characteristics discovery. We present IDAM, a novel method for disease-associated biomarker identification from metagenomic and metatranscriptomic data, without requiring prior metadata. It integrates gene context conservation and regulatory mechanism through a mathematical model for maximizing the number of connected components between local-low rank submatrices of a gene expression matrix and known uber-operon structures. We applied IDAM to 813 inflammatory bowel disease-associated datasets and showed IDAM outperformed existing methods in microbial biomarker identification. In addition, the identified biomarkers successfully distinguished disease subtypes and showcased their power in discovering novel disease subtypes/states. IDAM is freely available at https://github.com/OSU-BMBL/IDAM.

Keywords

Metagenome and metatranscriptome; Microbial biomarkers; Gene regulation; Gene context conservation; Causal inference; Unsupervised approach.

Background

Trillions of microbes colonize the human body and play critical roles in multiple fundamental physiological processes, such as immune development and dietary energy harvest [2]. Massive evidence reveals that changes in microbial composition and functions are intimately interwoven with multiple diseases ranging from obesity to cancer and autism [2, 3]. With the rapid development of sequencing technologies, worldwide projects, such as the Human Microbiome Project [4], the integrative Human Microbiome Project [5], and the Metagenomics of the Human Intestinal Tract project [6], have generated vast amounts of microbial data for diverse diseases. Thereafter, increasing efforts have been aimed to bring these data to clinical insights and deepen the understanding of crucial mechanisms responsible for disease progression and treatment. For example, recent studies have suggested that transferring several microbes from a donor into a sick recipient can promote the recipient's recovery and cause changes in the donor phenotype (e.g., increased adiposity) [7]. Furthermore, some pathogenic organisms have been found closely related to disease states and can be used as a potential signal of disease development [8, 9]. These findings hint at a promising avenue for clinical application (disease diagnosis and therapy) via the use of microbial key taxa or functions, i.e., microbial biomarkers.

Till now, a great deal of research has been explored for identifying microbial biomarkers due to their critical roles in diseases. One of the classical paradigms is identifying statistically significant taxa and functions that can differentiate two or more

groups as biomarkers [10-12]. Additionally, given machine learning's success under multiple scenarios, there are several attempts to adopt them in microbial marker identification, such as random forest and deep feedforward networks [13, 14]. However, these approaches face several issues.

First, most of them focus on 16S ribosome RNA (16S rRNA) and metagenomic (MG) data. 16S rRNA data suffer from a low taxonomic resolution and an absence of functional information [15]. The taxonomic analysis alone may induce spurious biomarkers since diverse microbial communities from different patients can perform remarkably similar functional capabilities [14]. While MG data can provide information at all taxonomic levels as well as potential functional information, researchers have found that: (i) a great proportion of reads cannot be successfully mapped to existing reference genomes during taxonomic identification, which leads to the potential loss of valuable information [16]. Intuitively, one may propose to use de novo assembled genomes as a compliment. However, the difficulties in metagenome assembly and binning generally cause incorrect contigs, which substantially affect further taxonomic inferences [17]. (ii) the functional analyses from metagenomics are not equivalent to the true activities within the microbiome. It has been observed that multiple metagenomically abundant genes in the human gut microbiome were significantly down-regulated at the transcriptional level [18].

Second, existing methods rely heavily on enough samples with detailed metadata information to design groups or training models. However, the current metadata of a considerable number of sequencing samples is incomplete, misleading, or not publically available [19], which may lead to these methods being infeasible or occurring bias in biomarker identification. Moreover, their intrinsic design in using known phenotype information makes them incapable of revealing new subtypes or stages of diseases. These challenges drive the need to develop an easy-to-use and effective method that keeps up with current data for disease-associated microbial biomarker identification.

In this study, we proposed a novel method to Infer Disease-Associated Microbial functional biomarkers (IDAM), which to a large extent surpasses these limitations discussed above. Specifically, we used matched MG and metatranscriptomic (MT) data (referred to sequencing data of DNA and RNA from the same sample) and integrated both gene regulation and gene context conservation during evolutionary to identify cross-species functional biomarkers without requiring metadata information. We applied IDAM to 813 published matched MG and MT datasets of inflammatory bowel diseases (IBD) (including two subtypes, ulcerative colitis (abbreviated as UC) and Crohn's disease (abbreviated as CD)) [20], and it could identify 41 gene modules as biomarkers. We validated the reliability of these biomarkers based on known phenotypes and well-studied species/functions related to IBD reported previously. Results suggested superior performance of IDAM compared to several popular tools, including LEfSe [12], ICA [21], QUBIC [22], ISA [23], and FABIA [24]. Furthermore, it distinguished two IBD subtypes well. We believe that IDAM is a highly advantageous method for biomarker inference based on

MG and MT data. It can pave the way to understanding the role of the microbiome in human diseases and guide disease diagnosis and therapy in clinical practice.

Results

IDAM: a metagenomic and metatranscriptomic analysis framework for disease-associated microbial biomarker identification

We first present the problem formulation of this article and then provide an overview of IDAM.

Problem formulation

The data we used is matched MG and MT data sequenced from microbial samples, giving a readout of active functions within communities. To better use available data, we focus on a community-level functional analysis that goes beyond species level (by which the reads do not need to be mapped to reference genomes first), providing a comprehensive overview of functions performed by either individuals or different microbial assemblages.

Since functionally related genes tend to show a coordinated expression pattern, one possible way for biomarker identification is to detect genes with distinguishable expression patterns under some subsets of samples. The gene sets can be inferred as biomarkers of the shared disease phenotype of the samples. However, co-expression is not equivalent to functional relevance. Furthermore, one concern is that gene expression referred here is assessed by the reads from various species (Fig.1a). Thus, it represents an ensemble expression of multiple homologous genes. One cannot fully characterize the complex functional activities based solely on the ensemble co-expression, by which the genes with no biological relevance may be clustered together (Fig.1b).

A previous study has suggested that functionally associated gene sets are generally highly conserved during evolution (i.e., the uber-operon structure) [25]. This finding strongly motivates us to integrate gene context conservation to study cross-species functional-related mechanisms, thereby alleviating the risk of false positives or false negatives. We use the uber-operon structure as the evolutionary footprints of cross-species functionally related genes. We believe genes are more likely to be functionally related if they have similar expression patterns and share the same uber-operon during evolution.

The above idea can be mathematically formulated as an optimization problem (Box 1 & Fig. 1c). However, this problem is theoretically intractable (NP-hard, Additional file 1: Section 1). Therefore, instead of solving it directly, we develop a heuristic algorithm, IDAM, for an approximate optimal solution.

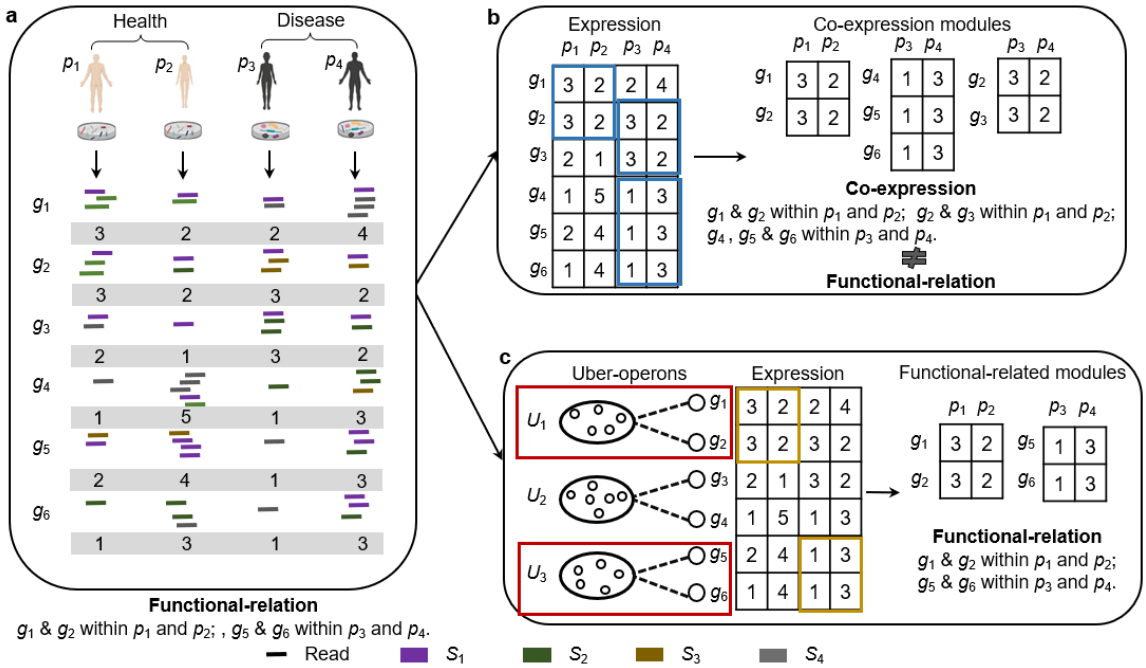


Fig.1 The diagram of problem formulation. **a** Sequencing reads from four samples (p_1 and p_2 with health phenotypes, and p_3 and p_4 with disease phenotype). Since metagenomic shotgun sequencing does not isolate microorganisms, the reads from different species (s_1, s_2, s_3, and s_4, represented by different colors) were jumbled together, and we cannot distinguish species-level information clearly. A ground truth behind these samples is that g_1 and g_2 are functional-related in healthy samples while g_5 and g_6 are in diseased samples. For each gene, we measured the expression based on the number of mapped reads, shown in each row. Note, the microbial community transcript number (metatranscriptomic data) will be corrected using the underlying genomic copy number (metagenome data) in a real case. Here, we simply showed the number of transcriptomic reads as expressions. **b** The co-expression modules. There were three co-expression modules by finding the gene and sample subsets with coordinated expression. But they were not equivalent to the known functional relation. **c** The functional modules by integrating gene context conservation and expression similarity. Each ellipse, including multiple genes, represents an uber-operon. We would like to find modules with co-expression and maximize the connected components with uber-operons. As a result, the two modules were found and formed two connected components (represented by red rectangles). In this way, we inferred functional relations between g_1 and g_2, as well as g_5 and g_6, which is consistent with the truth.

Box 1. The mathematical formulation of biomarker identification.

Input. A gene-sample expression matrix $A = (a_{ij})_{m \times n}$ from matched MG and MT data, which includes m genes ($g_1, g_2, \ldots, g_m$) and n samples ($p_1, p_2, \ldots, p_n$). a_{ij} ($1 \leq i \leq m$, $1 \leq j \leq n$) represents the expression of gene g_i within sample p_j, which is assessed by the number of transcript from metatranscriptomic data via metagenomic normalization.

Mathematical formulation. For matrix A, the consistent expression patterns present in a subset of genes and samples are equivalent to find local low-rank submatrices within A. Therefore, we can integrate it with a graph model to make the genes as likely as possible to share one uber-operon. Specifically, define a bipartite graph $G = (U, V, E)$, where each node $u \in U$ represents an uber-operon (all uber-operons are known based on a previous paper [1]), each node $v \in V$ represents *a* gene from A, and the edge $e \in E$ linked from a node $u \in U$ to a node $v \in V$ exists if the gene represented by v belongs to uber-operon u (Methods). Based on this graph and matrix A, the biomarker identification problem is transformed into identifying a set of significant local low-rank submatrices within A that maximizes the number of connected components between the corresponding gene sets and U on G. The largest number of connected components between gene subsets and U is to reduce the dispersion of genes in the uber-operons, thereby ensuring the genes within identified submatrices as functionally related as possible.

Output. Local low-rank submatrices within A, each of which contains a subset of genes and samples.

Overview of IDAM

The main framework of IDAM is used to seek disease-associated microbial biomarkers (Fig. 2). Firstly, the community-level expression within each sample is assessed by HUMAnN2 based on the matched MG and MT data [26]. The gene-sample expression matrix can be obtained, where each entry represents the expression value of a gene within a sample. Secondly, IDAM combines expression similarity and gene distribution within uber-operons to measure the possibility of each pair of genes that are involved in the same function, thereby obtaining a list based on the combined score for module initialization. Thirdly, the genes pair in the list is used as seeds one by one, and each seed is expanded by iteratively adding other functional-related genes to generate modules until the size of the new module goes down. Finally, the gene modules consisting of gene sets and corresponding sample sets will be output. We will regard the gene sets as functional biomarkers of the typical phenotype or state of the corresponding samples. Details are available in the “Methods” section.

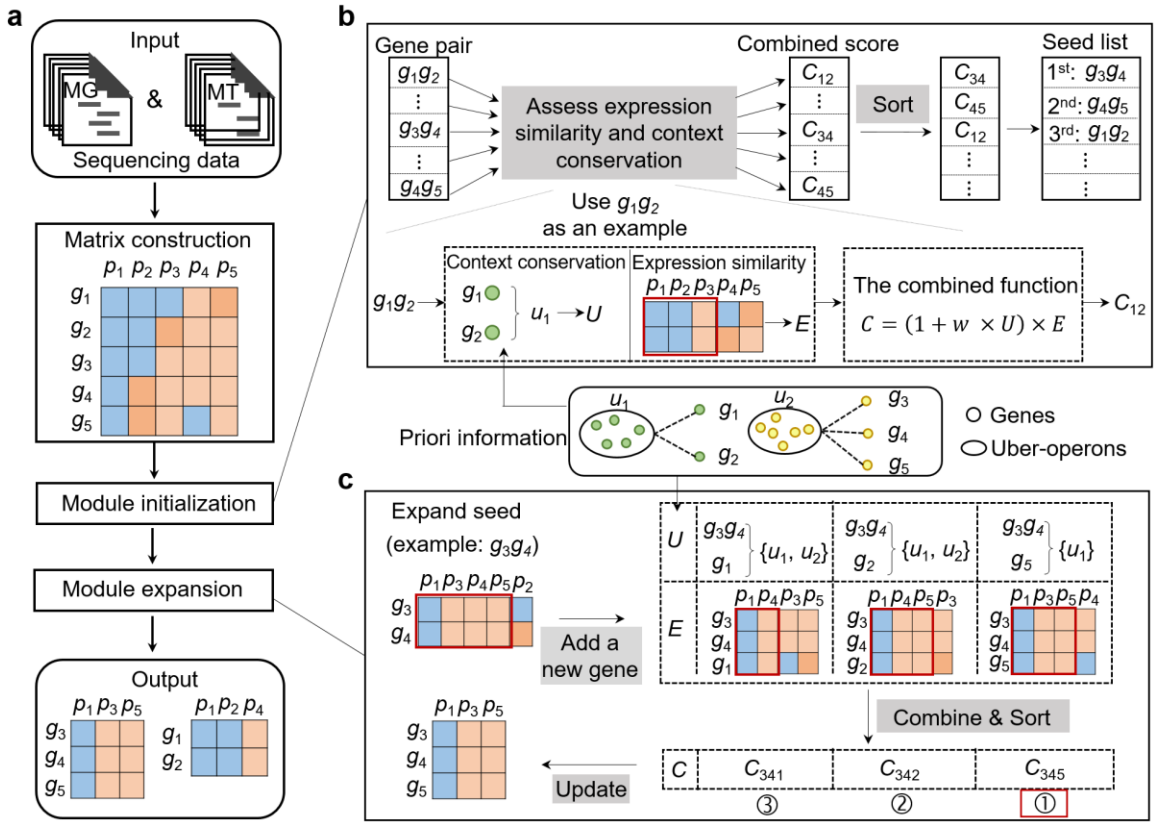


Fig.2 The workflow of IDAM. **a** The flow chart of IDAM. From the input matched MG and MT data of multiple samples (here are five), we can get an expression matrix (consisting of five samples). By module initialization and expansion, the modules will be output. **b** The details of module initialization. For each pair of genes, we assess expression similarity and context conservation, respectively, and get a combined score. The pair g_1g_2 is used as an example. Based on the priori knowledge of uber-operons, we know the two genes are from the same operon u_1. Therefore, we can assign a reward U for g_1g_2 (Methods) as the measurement of context conservation. Expression similarity is assessed according to the number of samples under which the rows are identical. There are three samples (p_1, p_2, and p_3) for g_1 and g_2, thus we can get a score E as expression similarity measurement (Methods). Finally, U and E are combined via a combined function C (w is a parameter, Methods), obtaining a score C_{12}. All gene pairs are sorted in decreasing order based on the combined scores. **c** The details of module expansion. We use the gene pair in the seed list as initial modules and expand them one by one. We iteratively add a new gene with the highest combined score into a module until the size of the updated module does down. Using the first pair g_3g_4 as an example. First, we add the three new genes (g_1, g_2, and g_5) into g_3g_4, respectively, forming three new modules ($g_3g_4g_1$, $g_3g_4g_2$, and $g_3g_4g_5$). Gene context conservation and expression similarity are respectively assessed (Methods). Since the module $g_3g_4g_5$ is of the highest score, we update the previous module (g_3g_4) as this one and the size is 3×3. Next, iteratively add a new gene (g_1 and g_2) to the current

module. Limited to space, we omit this. The processes are similar to the previous. After comparing the combined score, we will find $g_3g_4g_5g_2$ is of the highest score, and the new module is of size 4×2, which is smaller than the previous one (3×3) (Methods). Therefore, the module $g_3g_4g_5$ will be output.

The biomarkers identified by IDAM well-distinguished different phenotypes of cohorts and enabled us to uncover disease subtypes.

To assess the effectiveness of biomarkers from IDAM, we focused on whether identified biomarkers can distinguish different phenotypes. We applied IDAM to 813 publicly available datasets of patients with IBD and controls (referred to as non-IBD) (Methods, all following analyses are based on these datasets). By assessing the community-level expression within each sample, we obtained a matrix consisting of the expression of 941,785 genes within 813 samples. Based on this, IDAM found 41 gene modules with genes spanning from 2,340 to 41, totally covering 19,327 genes (Additional file 2: Table S2). This means there were 41 identified biomarkers (the gene set contained in each gene module is considered a biomarker). Ideally, each phenotype should have specific biomarkers, and samples with the same biomarker should be of highly consistent phenotypes. We used Z-test to evaluate the phenotype consistency of samples in each gene module and focused on the significantly occurred phenotype of the sample sets (Methods). Results showed that 32 biomarkers-associated sample sets were consistently enriched in a particular phenotype, and both IBD and Non-IBD were covered (19 enriched in IBD and 13 in Non-IBD) (Fig.3a, Additional file 2: Table S3). A noteworthy fact was that the identified biomarkers can recognize the subtypes of IBD, in which 11 correspond to UC and 8 correspond to CD (Fig.3b). We here gave an example of three modules (Module 7, Module 13, and Module 21 were selected, where each corresponds to a particular phenotype) to show the phenotype-specificity (Fig.3c). Limited by space, we showed the expression of the top 3% genes of each module (17/618, 19/689, and 14/521, respectively). Biological enrichment analysis (Methods) suggested these phenotype-specific modules are meaningful, and multiple enriched functions are paralleling with previous observations [27, 28]. For example, there was a significant enrichment of short-chain fatty acids metabolism (e.g., butyrate metabolism) in non-IBD samples relative to IBD. L-methionine biosynthesis, superpathway of L-citrulline metabolism, and superpathway of L-lysine, L-threonine, and L-methionine biosynthesis were enriched in Non-IBD individuals, while $preQ_0$ biosynthesis in IBD patients. Additionally, the superpathway of lipopolysaccharide biosynthesis was enriched in CD-associated biomarkers while methylphosphonate degradation in UC. These findings were consistent with previous studies [27, 29, 30], indicating that the biomarkers from IDAM were biologically informative. Collectively, IDAM has the capability to uncover biomarkers of disease phenotypes and can provide insights into the discovery of novel disease subtypes or states.

We compared IDAM with several popular biclustering approaches, including ICA [21], QUBIC [22], ISA [23], and FABIA [24] (Methods). These methods have not been applied to microbial data for biomarker identification. But intuitively, they can reveal functional characteristics as biomarkers under different phenotypes without the need for prior metadata. Results suggested that there were 34, 30, and 13 biomarkers found from QUBIC, ICA, and FABIA, respectively, while no results from ISA. Analogically, we focused on the phenotype consistency of biomarker-associated sample sets. While 78.05% (32/41) sample sets from IDAM were with a consistent phenotype, only 61.76% (21/34), 33.33%(10/30), and 30.77%(4/13) for QUBIC, ICA, and FABIA, respectively (Fig. 3d). This revealed that the inferred biomarkers from IDAM can best cluster the sample with the same phenotypes together compared with others, demonstrating the high effectiveness in distinguishing phenotypes. Indeed, this finding was not surprising since IDAM took both gene expression and gene context conservation into consideration, thereby identifying more biologically meaningful biomarkers, while others intrinsically used only gene expression. Detail case analyses can be found in the following subsection “The integration of gene context conservation substantially improves the performance of biomarker identification”. Additionally, we compared the running time of these methods. IDAM spent seven hours. By contrast, QUBIC and FABIA took 16 and 22 hours, while ISA was terminated after 24 hours without results. ICA costs eight minutes; however, the application power of ICA is limited by the relatively low accuracy (only one-third of the biomarkers can distinguish phenotypes).

In summary, the biomarkers identified by IDAM performed well in distinguishing phenotypes and can provide hints for new disease subtypes or states in clinical application.

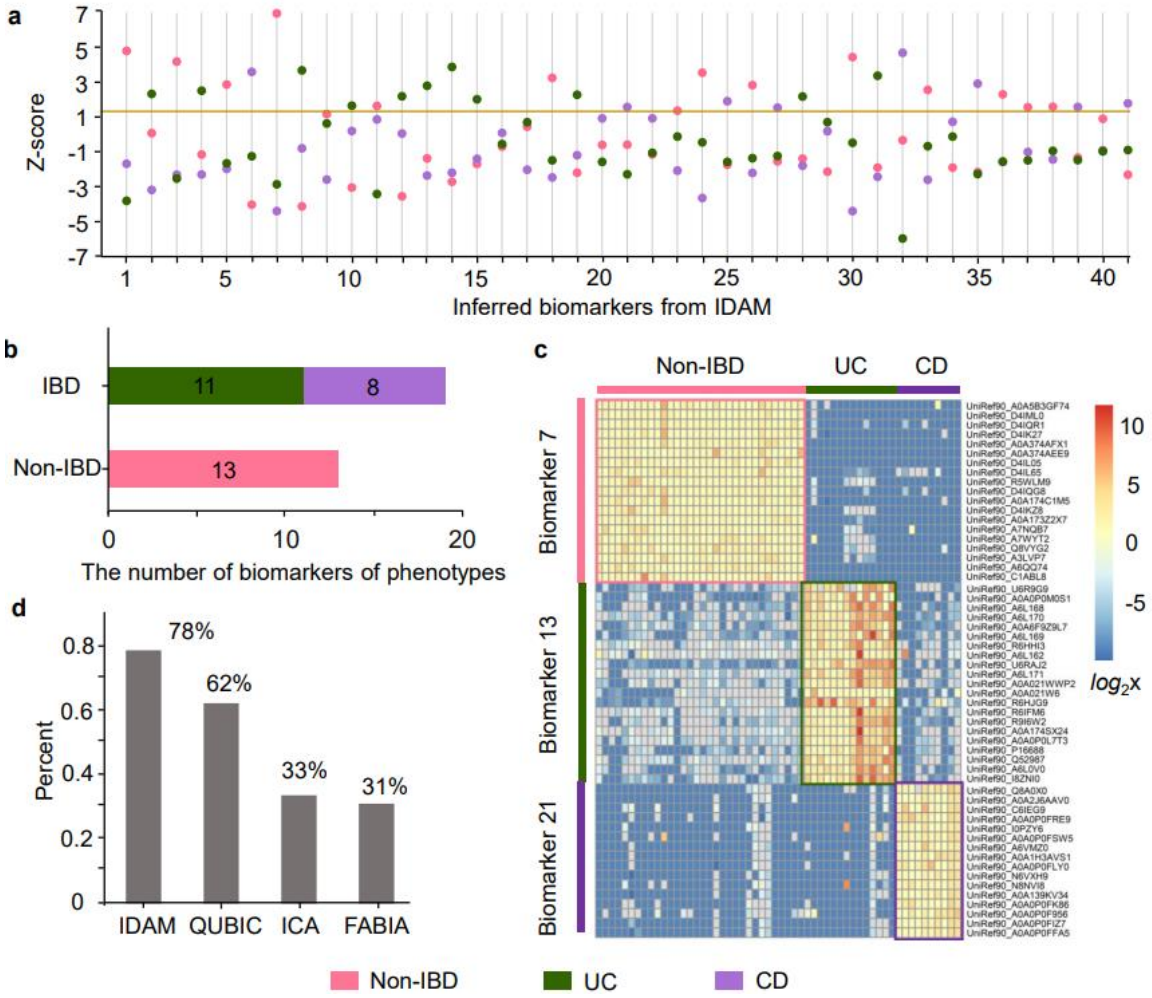


Fig.3 Observation on phenotypes of biomarker-associated sample sets. **a** The Z-score of Non-IBD, UC, and CD. They were respectively shown as pink, green, and purple, and each column represents a biomarker. The yellow line means the critical value of 1.645. The color of dots above this line means the corresponding phenotype exhibited strong enrichment within the biomarker-associated sample sets. **b** The number of biomarkers for each phenotype. A total of 19 biomarkers correspond to IBD (in which 11 to UC and 8 to CD) and 13 correspond to Non-IBD. **c** An example of gene modules associated with phenotypes (the log2 value of expression was showed). Module 7, 13, and 21 were used here. d The percent of biomarkers associated with samples of one phenotype.

Application of the causal inference model demonstrated high reliability of identified biomarkers.

We assessed whether the identified biomarkers from IDAM had true associations with phenotypes. Since microbial communities are influenced by multiple factors such as age, race, and diet, they are of high inter-individual heterogeneity, prone to overwhelm disease-

associated microbial characteristics, and increase the risk of false positives or negatives in biomarker studies. We used a causal inference model for each biomarker to test association with phenotypes. We controlled for other covariates, such as age, sex, and comorbidities, and constructed paired samples (Methods). Wilcoxon signed-rank test based on the paired samples indicated that 82.93% (34 of 41) biomarkers were significantly related to a specific phenotype (with *p*-value < 0.05, Fig.4, Additional file 2: Table S4), demonstrating high reliability of identified biomarkers from IDAM.

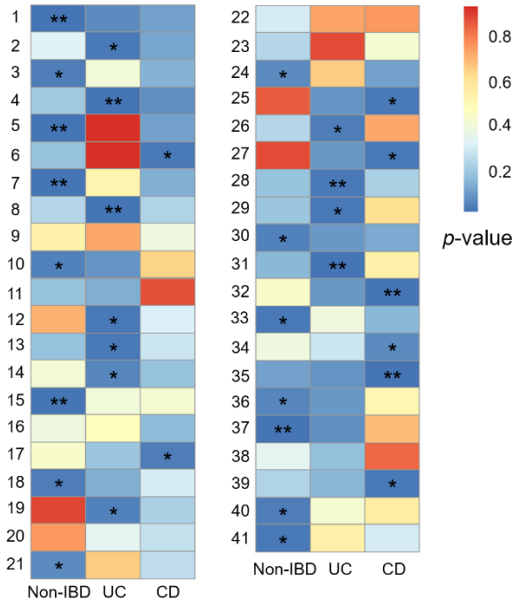


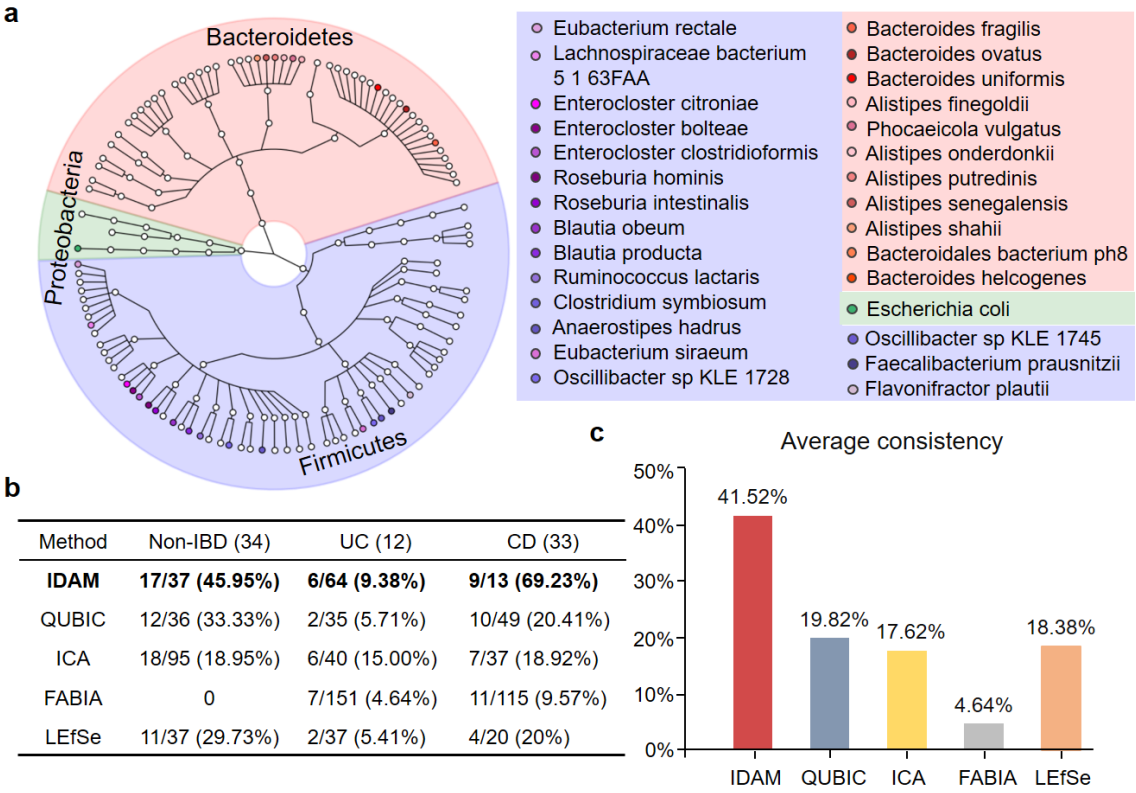
Fig.4 The result of biomarkers from the Wilcoxon rank-sum test. Each row represents a biomarker, and the color means *p*-value. *p-value* is indicated as follows: *, <0.05; **, <0.01.

Species-level analysis suggested the decent performance of IDAM

We implemented species-level analysis for biomarkers from IDAM. Since differences in community-level functions ultimately imply species-level functional differences, we decomposed genes within each biomarker into species-specific data (Methods). As a result, these genes were from 111 species, which were dominated by Bacteroidetes and Firmicutes, followed by Proteobacteria at the phylum level (Fig. 5a). This observation was consistent with previous research about IBD-associated species [28]. We further classified the species according to biomarker-associated phenotypes (Methods) and obtained 37, 64, and 13 species associated with Non-IBD, UC, and CD, respectively (Additional file 2: Table S5). Three of them were overlapped between UC and CD, including *Flavonifractor plautii*, *Bacteroides vulgatus*, and *Bacteroides fragilis*. We collected IBD-related species (34 for Non-IBD, 12 for UC, and 33 for CD) from previous taxonomic studies [31-33] (Additional file 2: Table S6). Among them, a total of 29 species were captured by IDAM

(Fig. 5a), including 17 for Non-IBD, 6 for UC, and 9 for CD (red in Additional file 2: Table S5). An interesting finding was that the three overlapped species from IDAM have been reported associated with both UC and CD [31-33], suggesting the rationality of IDAM results.

We compared IDAM with ICA [18], QUBIC [19], FABIA [21], and LEfSe [12] (Methods), in which the first three are not dependent on metadata while LEfSe identifies differential abundant species as biomarkers based on prior phenotype information. Detailed results are showed in Fig.5b. Not surprisingly, the average consistency with collected species across three phenotypes was highest for IDAM (41.52%), followed by QUBIC (19.82%), LEfSe (18.38%), ICA (17.62%), and FABIA (4.74%) (Methods, Fig. 5c). Collectively, the species-level analysis showed a decent performance of IDAM.



Method	Non-IBD (34)	UC (12)	CD (33)
IDAM	**17/37 (45.95%)**	**6/64 (9.38%)**	**9/13 (69.23%)**
QUBIC	12/36 (33.33%)	2/35 (5.71%)	10/49 (20.41%)
ICA	18/95 (18.95%)	6/40 (15.00%)	7/37 (18.92%)
FABIA	0	7/151 (4.64%)	11/115 (9.57%)
LEfSe	11/37 (29.73%)	2/37 (5.41%)	4/20 (20%)

Fig.5 The species-level analysis of identified biomarkers. **a** The phylogenetic tree of the species associated with identified biomarkers from IDAM. The colored dots indicated the species that have been validated by previous studies. **b** The results of species analysis by different methods. The numbers in the header (34, 12, and 33) represent the number of collected species associated with the three phenotypes, respectively. The number behind the slash represents the number of species associated with a particular phenotype based on different methods, among which consistent with the collected species are represented before the slash. Percentages indicate the consistency index of each method associated with phenotypes. **c** The mean of consistency under three phenotypes was

used as the average consistency for each method. The greater value means identified biomarkers-associated species are more reliable.

The integration of gene context conservation substantially improved the performance of biomarker identification.

We investigated the contribution of gene context conservation, given our underlying hypothesis that genes with similar expression patterns and high context conservation during evolution are more likely to be functionally related. Therefore, we seek to identify biomarkers without combining gene context conservation. Observation showed that the performance was significantly decreased compared with IDAM. Overall, the pathway enrichment analysis suggested that 67.65% (23 of 34) biomarkers from IDAM were significantly enriched with at least one MetaCyc functional category, while 82.93% (34/41) of IDAM. This preliminarily indicated that the integration of the uber-operon structure contributed to clustering genes with biological relevance together, which improved the ability of IDAM to identify meaningful biomarkers.

We further explored the functional association between uber-operons and disease biomarkers. For each biomarker, we extracted underlying uber-operons based on the alignment relationships with included genes of the biomarker. Results showed that the enriched functions of biomarkers were significantly associated with underlying uber-operons. Take the purine metabolism pathway as an example. Previous studies have shown that the imbalance between biosynthesis and degradation of purine metabolism could produce excessive uric acid in the gut, leading to the occurrence of IBD [34]. As expected, the pathway was enriched in an IBD-associated biomarker of IDAM. We specifically observed the dominant uber-operon of the biomarker and found its primary function was purine metabolism [1]. The functional consistency between identified biomarkers and uber-operons indicated the strong power of gene context conservation in biomarker identification.

Discussion

Although many studies have focused on the microbial biomarker identification of diseases, it is still a fundamental challenge to explore critical characteristics in large-scale and highly heterogeneous microbial data. We developed a new methodology, IDAM, for matched MG and MT data to identify disease-associated biomarkers. Unlike the previous microbial biomarker analysis methods, a key innovation is that we incorporate gene context conservation and regulation information by leveraging a multi-omic view of microbial communities for biological features characterization of diseases. Furthermore, IDAM does not require any prior knowledge about samples, which alleviates bias from misleading data noise and allows us to reveal novel disease subtypes or states. We tested IDAM in IBD biomarker identification and showed that it remarkably outperformed existing methods.

We discovered 32 significant microbial biomarkers with IBD, and some of them are consistent with what has been reported previously. Specifically, we compared the method that removes gene context conservation information from IDAM and found that the biological relevance was significantly decreased between genes within a module, which demonstrated that the uber-operon structure played a crucial role in functional indication. Notably, though we only tested IDAM in IBD-related analysis, IDMA is ready to contribute to the understanding of any diseases.

Admittedly, some limitations can be found in our method. One potential issue of IDAM is that the provided biomarkers are at the community level, which needs further taxonomic analysis to improve their utility as clinical biomarkers. One feasible method is tracing the community-level functions back to the taxonomic level by HUMAnN2. However, there were some biomarkers that attributed to unknown species (labeled as unclassified species in HUMAnN2). Therefore, cautious inference is needed while interpreting these biomarkers since we cannot precisely distinguish whether they are false biomarkers or just from species that have not been well studied. However, this issue may not be problematic as more microbes are discovered and explored. Hence, we will continue to develop a three-dimensional analysis approach, taking microbial species and genes (pathways) of multiple samples into consideration for biomarker identification in the future. In this way, the specific modules with diseases consisting of function (genes/pathways) and species will be detected, simultaneously providing functional and taxonomic insights into diseases.

The second issue is that the need for large-scale samples. Since IDAM used expression similarity to assess the relevance of gene regulation, more samples will contribute to the reliable measurement from data. One can integrate the 813 provided samples to increase the sample pool, thereby guaranteeing the practicability of IDAM.

Additionally, the provided biomarkers of IDAM only provided microbiome-disease associations. Therefore, rigorous experiments are required to test causality between the identified biomarkers and the disease of interest and establish clinical prevalence and utility.

Conclusions

IDAM provides a highly effective approach to identify disease-associated microbial biomarkers based on gene regulation and context conservation during evolution, based on MG and MT data. The identified biomarkers of diseases can be useful in understanding disease pathogenesis and providing guidance for future development of disease diagnosis and treatment.

Methods

Datasets

We downloaded all matched quality-controlled MG and MT data, as well as the corresponding HUMAnN2 results and metadata at the IBDMDB website in March, 2020 (https://ibdmdb.org) [20]. In total, there are 198 samples from non-IBD subjects and 615 samples from patients with IBD (consisting of 381 UC samples and 234 CD samples). The uber-operon structure was from the paper of Che *et al.* [1].

IDAM: a framework of identifying disease-associated microbial biomarkers based on metagenomic and metatranscriptomic data

Step 1*: *Extracting community-level gene expression of samples.

The input data is quality-controlled (including quality and length filtered and screened for residual host DNA) matched MG and MT sequencing datasets (fasta, fasta.gz, fastq, or fastq.gz). Each matched dataset is firstly processed using HUMAnN2 for gene relative expression within communities [26] (In this study, we directly used the HUMAnN2 results of the 813 datasets from IBDMDB). Then, we extract the community-level gene expression of each sample to construct a gene-sample expression matrix $A_{m\times n}$, in which each row represents a gene (indicated by UniRef90 gene family [35]), and each column represents a sample.

***Step 2*: *Measuring gene context conservation and expression similarity for each gene pair*.**

We assess context conservation and expression similarity for each pair of genes. Given the large scale of genes in matrix $A_{m\times n}$ (the gene list is a pan-genome of all microorganisms from multiple samples), we firstly use the technique developed in previous studies to calculate an integer *t* and partition the rows within the matrix equally into *t* subsets [36, 37]. The aim is to reduce computation burden without affecting final results (i.e., keep all the significant submatrices).

Specifically, a low-rank submatrix is regarded as significant if it cannot occur randomly in matrix $A_{m\times n}$. Intuitively, the smaller the submatrix is, the more likely it is to appear randomly. To make the identified submatrices meaningful, we first statistically assess the size of the smallest submatrix that is impossible to occur randomly. All identified matrices not smaller than this will be thought to be significant. Suppose a local low-rank submatrix is of size $k \times s$, its possibility occurred in $A_{m\times n}$ is as follows [36, 37].

$$P_{ks} = n \times (n-1) \times \cdots \times (n-s+1) \times \binom{m}{k} \times \left(\frac{2}{s!}\right)^k \times (1-\frac{2}{s!})^{m-k}$$

We here think the submatrices is significant if P_{ks} < 0.05. When users have a pre-estimation of *s*, we can calculate *k* based on the formula mentioned above. By this, the minimal size of significant submatrices can be determined. Note that the submatrices from our algorithm are generated by a gene pair (Step 3). To keep all the important submatrices, we need to make sure at least one gene pair within each submatrices exists as a seed. Thus, we can at most divide the rows of matrix A into *k* -1 groups (i.e., *t* = *k* -1). Based on the above, users can set the number of groups. For IDAM, the default

value of t is 10. That is, the matrix will be divided into 10 groups. which can greatly reduce computation burden without affecting final results.

2.1 Gene context conservation

Gene context conservation is assessed based on gene distribution in uber-operons. Since the genes within the expression matrix referred here are gene families in UniRef90 (a non-redundant protein sequence database) [35], we align them with genes in uber-operons to determine which uber-operons these gene families belong to. Specifically, the sequences of genes within all uber-operons are collected as a database. For each gene family within the matrix, we extract the corresponding sequence from the UniProt database [35] and align it with the constructed database using Blast [38]. We think a gene family belongs to an uber-operon if the *E*-value of its alignment with one of the uber-operon's sequences is less than 0.001 (If multiple sequences from different uber-operons satisfy this, we will choose the smallest *E*-value to identify the uber-operon to which it belongs).

For each gene pair (g_i, g_j) within a partitioned subset (where g_i and g_j respectively indicate i th and j th gene within a subset with q genes, $1 \leq i < j \leq q$), we set a reward $U(g_i, g_j)$ as context conservation measurement based on their relative positions in uber-operons.

$$U(g_i, g_j) = \begin{cases} 1, & if\ g_i\ and\ g_j\ are\ within\ one\ uber-operon, \\ -1, & if\ g_i\ or\ g_j\ is\ included\ in\ one\ uber-operon\ and\ the\ other\ is\ not, \\ -2, & if\ g_i\ and\ g_j\ are\ included\ in\ different\ uber-operons, \\ 0, & otherwise. \end{cases}$$

2.2 Expression similarity of genes

For ease of measuring the similarity of expression, the continuous expression value of each gene is discretized to an integer representation by qualitative representation [22] (Additional file 1: Section 2). The expression similarity $E(g_i, g_j)$ of gene pair (g_i, g_j) is set as the number of samples under which the corresponding integers along the rows of the two genes are identical.

2.3 The integration of gene context conservation and expression similarity

A combined score $C(g_i, g_j)$ integrating gene context conservation and expression similarity of g_i and g_j is obtained via the following function,

$$C(g_i, g_j) = \left(1 + w \times U(g_i, g_j)\right) \times E(g_i, g_j)$$

where w is a parameter that can be set by users, ranging from 0 to 1, and a larger value of w indicates gene context conservation has a stronger impact on estimating functional relations. Here we set different values of w and then selected the one (the value 0.1 was selected) with the best performance as the default value (Additional file 1: Section 3 and Additional file 2: Table S7). Finally, the gene pairs are sorted in descending order for the combined scores, generating a list *L*.

Step 3*: *Generating modules consisting of gene sets (as biomarkers) and corresponding sample sets

3.1 Initialization

We start with the first gene pair in *L* that satisfies i) at least one gene of these two has not been included in previous modules, or ii) these two are respectively included within different previous modules, and there is no overlap between the gene sets of the two modules. We aggregate samples under which the integer of the two genes are identical (represented as set *S*), forming current module *M* = {*G*, *S*}, where G represents the gene set consisting of the paired genes. Then, the gene pair is removed from *L*.

3.2 Expansion

We expand *M* by adding a new gene (if any) that has the highest combined score with *M,* giving rise to the updated module *M'* = {*G'*, *S'*}. The calculation of the combined score between a gene (represented as g_z ($1 \leq z \leq m$)) and module *M* is similar to that of two genes g_i and g_j. Specifically, we first assess the context conservation with g_z and module *M*. We construct an uber-operon set *B* based on the uber-operons that the included genes of *M* belong to. A reward $U(g_z, M)$ as context conservation of g_z and *M* is set as follows.

$$U(g_p, M) = \begin{cases} 1, & if\ g_z\ is\ within\ the\ uber-operon\ set\ B, \\ -1, & if\ g_z\ is\ within\ uber-operons\ that\ have\ not\ included\ in\ B\ and\ uber-operon\ sets\ of\ previously \\ -2, & if\ g_z\ is\ included\ in\ an\ uber-operon\ that\ is\ outside\ of\ B, \\ 0, & otherwise. \end{cases}$$

Then, we define the expression similarity of g_z and *M* as the number of samples within *S* under which the corresponding integers along the rows are identical (or identical integers but with opposite signs). Based on these, a combined score $C(g_z, M)$ can be obtained by replacing $U(g_i, g_j)$ and $E(g_i, g_j)$ in the calculation of $C(g_i, g_j)$ to $U(g_z, M)$ and $E(g_z, M)$, respectively.

If min{|*G'*|, |*S'*|} > min{|*G*|, |*S*|}, we will set *M* = *M'* and repeat Step 3.2. Otherwise, go to step 4.

***Step 4*:** *Output*

Module *M* will be output if the proportion of overlapped genes and samples between *G* and the gene sets of previously identified modules is less than a threshold *f* (default f is 0.1 for more separate modules). Go to step 3 until the list *L* is empty.

Evaluating the phenotype consistency of samples within one gene module

Z-test was used to assess whether samples with the same biomarker were of highly consistent phenotypes. Specifically, we extracted the phenotypes of all samples from metadata as background and the corresponding phenotypes of samples in each gene

module as test sets. For each phenotype (including Non-IBD, UC, and CD) and each test set, we counted the phenotype occurrence in the test set and background, respectively. We evaluated the significance of each phenotype using a right-tailed Z-test with the null hypothesis that the observed frequency in the test set was less than or equal to the occurrence probability within the background. The threshold of significant level 0.05 was used with a critical Z-value of 1.645. If Z-score was greater than 1.645 (*p*-value > 0.05), the null hypothesis would be rejected, meaning the phenotype significantly occurred within the test set. We would consider the samples within the gene module was of a highly consistent phenotype.

Pathway enrichment analysis: UniRef90 gene families to pathway mapping

All UniRef90 identifiers were firstly mapped to MetaCyc reactions [26, 39]. The reactions were further associated with MetaCyc pathways [26, 39], which enabled a transitive association between UniRef90 gene families and MetaCyc pathways. We defined a pathway that was satisfied when the reactions within the pathway significantly occurred via the hypergeometric test,

$$P(t_{ij}) = \frac{\binom{n_j}{t}\binom{N-n_j}{m_i-t}}{\binom{N}{m_i}}$$

$$p-value = \sum_{t_{ij}>T_{ij}}^{n_j} P(t_{ij})$$

where N is the total number of reactions that associate with UniRef90 gene families and pathways in MetaCyc, m_i is the number of reactions within the i th biomarkers, n_j is the total number of reactions within the j th pathway, and T_{ij} is the observed number of reactions of pathway j occurred in the biomarker i. This yielded a p-value for each biomarker and pathway. We selected the pathways with a p-value less than 0.05 as the significantly enriched pathways of each biomarker.

Comparing IDAM with other methods

Biomarker identification of IDAM was compared with the other four top performed unsupervised tools for grouping genes into functional modules [40], including biclustering-based (QUBIC [22], ISA [23], and FABIA [24]) and decomposition-based (fastICA, an improved method of ICA [21]). All tools were tested and compared based on the filtered gene-sample matrix with 941,785 genes. We run these tools on the Pitzer cluster of the Ohio Supercomputer Center with memory usage set to 300GB [41] under default

parameters. Specifically, QUBIC was performed using the published source code, while the other three were implemented by the R packages isa2 0.3.5, fabia 2.36.0, and fastICA 1.2.2, respectively.

The causal inference model

Our aim was to assess whether the association between identified biomarkers and disease phenotypes was reliable. For each biomarker, we firstly estimated underlying confounding variables that affected its occurrence. Here, we incorporated diet, age, race, health conditions of other diseases, as well as the other biomarkers and calculated a propensity score for each sample (using R package MatchIt 4.2.0). Then, the paired samples were constructed by matching with the optimal score (by setting the parameter method = “optimal”). Finally, Wilcoxon signed-rank test was used to measure the significance of the biomarker with phenotypes, with a p-value < 0.05 as significantly associated.

Species-level analysis

For biomarkers from IDAM, ICA [18], QUBIC [19], and FABIA [21], we extracted species that contribute to the abundance of genes within identified biomarkers, respectively, based on HUMAnN2 output [26]. By this, community-level gene expression (abundance) was decomposed into species-level data. Since the biomarkers have been associated with phenotypes by Z-test mentioned above, we assigned the species of each biomarker to the corresponding phenotype of the biomarker. For each phenotype, we counted the total number of species, as well as the number that matched with the collected species. Then, the total number of species in the phenotype dividing by the number of matched species was calculated as the consistency index.

Besides, another widely used tool, LEfSe [12], was performed here for comparison. It was based on species profiling from MetaPhlAn2 to identify statistically different species among the three phenotypes [42, 43]. The significantly differential species of each phenotype were aligned to collected ones, by which we determined the number of matched species and calculated the consistency of each phenotype.

Abbreviations

16S ribosome RNA: 16S rRNA;

Metagenomic: MG;

Metatranscriptomic: MT;

Inflammatory bowel disease: IBD;

Ulcerative colitis: UC;

Crohn's disease: CD.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets supporting the conclusions of this article are available in the IBDMDB, https://ibdmdb.org.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Key R&D Program of China (2020YFA0712400), National Nature Science Foundation of China (NSFC, 61772313 and 11931008), and Interdisciplinary Science Innovation Group Project of Shandong University (2019).

Authors' contributions

J.Z., and B.L. conceived the basic idea. Z.L. carried out the computational analysis and data interpretation. L.Z. and Q.W. designed and drew the figures. D.C. and Q.M. helped polished the manuscript and all the authors wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Anjun Ma for his assistance in the design of Fig. 1 and Fig.2.

References

1. Che D, Li G, Mao F, Wu H, Xu Y. Detecting uber-operons in prokaryotic genomes. Nucleic acids research. 2006; 34**:**2418-2427.

2. Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science. 2021; 371.
3. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature medicine. 2018; 24:392-400.
4. Consortium THMP. A framework for human microbiome research. Nature. 2012; 486:215-221.
5. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, Buck GA, Snyder MP, Strauss JF, Weinstock GM, et al. The Integrative Human Microbiome Project. Nature. 2019; 569:641-648.
6. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59-65.
7. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006; 444:1027-1031.
8. Dubinsky M, Braun J. Diagnostic and Prognostic Microbial Biomarkers in Inflammatory Bowel Diseases. Gastroenterology. 2015; 149:1265-1274.e1263.
9. Lecuit M, Lortholary O. Immunoproliferative small intestinal disease associated with Campylobacter jejuni. Med Mal Infect. 2005; 35:S56-58.
10. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013; 10:1200-1202.
11. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. Briefings in Bioinformatics. 2020; 22:1639-1655.
12. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biology. 2011; 12:R60.
13. Zhu Q, Jiang X, Zhu Q, Pan M, He T. Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. Frontiers in Genetics. 2019; 10.
14. Wu Y, Jiao N, Zhu R, Zhang Y, Wu D, Wang A-J, Fang S, Tao L, Li Y, Cheng S, et al. Identification of microbial markers across populations in early detection of colorectal cancer. Nature Communications. 2021; 12:3063.
15. Snipen L, Angell I-L, Rognes T, Rudi K. Reduced metagenome sequencing for strain-resolution taxonomic profiles. Microbiome. 2021; 9:79.
16. Zhu Z, Ren J, Michail S, Sun F. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biology. 2019; 20:154.
17. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology. 2021; 39:105-114.
18. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, et al. Relating the metatranscriptome and metagenome of the human gut. Proceedings of the National Academy of Sciences. 2014; 111:E2329-E2338.

19. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, Jehmlich N, Herberth G, von Bergen M, Stadler PF, Carvalho André Carlos Ponce de Leon Ferreira d, Nunes da Rocha U. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Research. 2020; 49:D743-D750.
20. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569:655-662.
21. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Networks. 2000; 13:411-430.
22. Li G, Ma Q, Tang H, Paterson AH, Xu Y. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Research. 2009; 37:e101-e101.
23. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys. 2003; 67:031902.
24. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Van Sanden S, Lin D, Talloen W, et al. FABIA: factor analysis for bicluster acquisition. Bioinformatics. 2010; 26:1520-1527.
25. Lathe III WC, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends in biochemical sciences. 2000; 25:474-479.
26. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods. 2018; 15:962-968.
27. Imhann F, Vich Vila A, Bonder MJ, Fu J, Gevers D, Visschedijk MC, Spekhorst LM, Alberts R, Franke L, van Dullemen HM, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut. 2018; 67:108-119.
28. Schirmer M, Garner A, Vlamakis H, Xavier RJ. Microbial genes and pathways in inflammatory bowel disease. Nature reviews Microbiology. 2019; 17:497-511.
29. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology. 2012; 13:R79.
30. Klaassen MAY, Imhann F, Collij V, Fu J, Wijmenga C, Zhernakova A, Dijkstra G, Festen EAM, Gacesa R, Vich Vila A, Weersma RK. Anti-inflammatory Gut Microbial Pathways Are Decreased During Crohn’s Disease Exacerbations. Journal of Crohn's and Colitis. 2019; 13:1439-1449.
31. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature microbiology. 2019; 4:293-305.

32. Kuballa A, Geraci M, Johnston M, Sorrentino D. The Gut Microbial Profile of Preclinical Crohn's Disease Is Similar to That of Healthy Controls. Inflamm Bowel Dis. 2020; 26:1682-1690.
33. Vich Vila A, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z, Kurilshikov A, Bonder MJ, Jiang X, Tigchelaar EF, et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci Transl Med. 2018; 10.
34. Wu J, Wei Z, Cheng P, Qian C, Xu F, Yang Y, Wang A, Chen W, Sun Z, Lu Y. Rhein modulates host purine metabolism in intestine through gut microbiota and ameliorates experimental colitis. Theranostics. 2020; 10:10665-10679.
35. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23:1282-1288.
36. Wang Z, Li G, Robinson RW, Huang X. UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data. Sci Rep. 2016; 6:23466.
37. Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. J Comput Biol. 2003; 10:373-384.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403-410.
39. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 2020; 48:D445-d453.
40. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nature Communications. 2018; 9:1090.
41. Center OS. Ohio Supercomputer Center. 1987.
42. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012; 9:811-814.
43. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015; 12:902-903.

Supplementary Information

Additional file 1: Supplementary material

This supplementary material includes details in the Problem formulation and Methods part. Section 1 illustrates why the problem we formulated is NP-hard. Sections 2 and 3 describe how to do discretization and select parameter *w*.

Additional file 2: Supplementary results

This file includes supplementary results in the Result part and parameter selection of *w* in the Method part. Table S1 is an overview of Additional file 2. Table S2 shows the size of identified gene modules from IDAM. Table S3 shows the Z-score and corresponding *p*-

value of each module for three phenotypes. Table S4 shows the *p*-value from the causal inference model. Table S5 shows the species associated with different phenotypes from IDAM. Table S6 shows the species associated with different phenotypes collected from previous studies. Table S7 shows the performance of IDAM with different parameter *w*, which is attached to Methods and Additional file 1: Section 3.